



## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

<b>(51) International Patent Classification <sup>6</sup>:</b> <b>C07H 21/00, 21/04</b>	<b>A1</b>	<b>(11) International Publication Number:</b> <b>WO 97/11086</b> <b>(43) International Publication Date:</b> 27 March 1997 (27.03.97)						
<b>(21) International Application Number:</b> PCT/US96/15088 <b>(22) International Filing Date:</b> 20 September 1996 (20.09.96) <b>(30) Priority Data:</b> 08/532,390 22 September 1995 (22.09.95) US <b>(71) Applicant:</b> THE GENERAL HOSPITAL CORPORATION [US/US]; 55 Fruit Street, Boston, MA 02114 (US). <b>(72) Inventors:</b> SEED, Brian; Apartment 5J, Nine Hawthorn Place, Boston, MA 02114 (US). HAAS, Jurgen; Huberweg 13, D- 69198 Schriesheim (DE). <b>(74) Agent:</b> LECH, Karen, F.; Fish & Richardson P.C., 225 Franklin Street, Boston, MA 02110-2804 (US).		<b>(81) Designated States:</b> CA, JP, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).  <b>Published</b> <i>With international search report.</i> <i>Before the expiration of the time limit for amending the          claims and to be republished in the event of the receipt of          amendments.</i>						
<b>(54) Title:</b> HIGH LEVEL EXPRESSION OF PROTEINS  <b>(57) Abstract</b> <p>The invention features a synthetic gene encoding a protein normally expressed in a mammalian cell or eukaryotic cell wherein at least one non-preferred or less preferred codon in the natural gene encoding the mammalian protein has been replaced by a preferred codon encoding the same amino acid.</p> <table border="0"> <tr> <td style="vertical-align: top;">           1 51 101 151 201 251 301 351 401 451 501 551 601 651 701 751         </td> <td style="vertical-align: top;">           GAATTCACGC GTTACACGGG GCCACAAGTT AAGCTGACCC GCCCCACCCTC ACCCCGACCA GGCTACGTCC GACCCGCGCC AGCTGAAGGG CTGGAGTACA GAAGAACGGC GCAGCGTGCA GGCCCCGTGC GAGCAAAGAC TGACCGCCGC GGCCGCGGAT         </td> <td style="vertical-align: top;">           GTAAGCTTGC GTGGTGCCCA CAGCGTGTC TGAAGTTCAT GTGACCACCT CATGAAGCAG AGGAGCGCAC GAGGTGAAGT CATCGACTTC ATCAAGGTGA GCTCGCCGAC TGCTGCCCCG CCCAACGAGA CGGGATCACT CC         </td> <td style="vertical-align: top;">           CGCCACCATG TCCTGGTCGA GGCGAGGGCG CTGCACCACC TCAGCTACGG CACGACTTCT CATCTTCTTC TCGAGGGCGA AAGGAGGACG GCAACATCCT CACTACCAGC CAACCACTAC AGCGCGATCA CACGGCATGG         </td> <td style="vertical-align: top;">           GTGAGCAAGG GCTGGACGGC AGGGCGATGC GGCAAGCTGC CGTGCACTGC TCAAGTCCGC AAGGACGACG GCAACATCCT CCGCCACAAC AGAACACCCC CTGAGCACCC CAACCACTAC ACGAGCTGTA         </td> <td style="vertical-align: top;">           GCGAGGAGCT GACGTGAACG CACCTACGGC CCGTGCCCTG TTCAGCCGCT CATGCCCGAA GCAACTACAA AACCGCATCG GGGGCACAAG CCGACAAGCA ATCGAGGACG CATCGGCGAC AGTCCGCCCT CTGGAGTTCTG CAAGTAAAGC         </td> </tr> </table> <p style="text-align: right;">(SEQ ID NO: 40)</p>			1 51 101 151 201 251 301 351 401 451 501 551 601 651 701 751	GAATTCACGC GTTACACGGG GCCACAAGTT AAGCTGACCC GCCCCACCCTC ACCCCGACCA GGCTACGTCC GACCCGCGCC AGCTGAAGGG CTGGAGTACA GAAGAACGGC GCAGCGTGCA GGCCCCGTGC GAGCAAAGAC TGACCGCCGC GGCCGCGGAT	GTAAGCTTGC GTGGTGCCCA CAGCGTGTC TGAAGTTCAT GTGACCACCT CATGAAGCAG AGGAGCGCAC GAGGTGAAGT CATCGACTTC ATCAAGGTGA GCTCGCCGAC TGCTGCCCCG CCCAACGAGA CGGGATCACT CC	CGCCACCATG TCCTGGTCGA GGCGAGGGCG CTGCACCACC TCAGCTACGG CACGACTTCT CATCTTCTTC TCGAGGGCGA AAGGAGGACG GCAACATCCT CACTACCAGC CAACCACTAC AGCGCGATCA CACGGCATGG	GTGAGCAAGG GCTGGACGGC AGGGCGATGC GGCAAGCTGC CGTGCACTGC TCAAGTCCGC AAGGACGACG GCAACATCCT CCGCCACAAC AGAACACCCC CTGAGCACCC CAACCACTAC ACGAGCTGTA	GCGAGGAGCT GACGTGAACG CACCTACGGC CCGTGCCCTG TTCAGCCGCT CATGCCCGAA GCAACTACAA AACCGCATCG GGGGCACAAG CCGACAAGCA ATCGAGGACG CATCGGCGAC AGTCCGCCCT CTGGAGTTCTG CAAGTAAAGC
1 51 101 151 201 251 301 351 401 451 501 551 601 651 701 751	GAATTCACGC GTTACACGGG GCCACAAGTT AAGCTGACCC GCCCCACCCTC ACCCCGACCA GGCTACGTCC GACCCGCGCC AGCTGAAGGG CTGGAGTACA GAAGAACGGC GCAGCGTGCA GGCCCCGTGC GAGCAAAGAC TGACCGCCGC GGCCGCGGAT	GTAAGCTTGC GTGGTGCCCA CAGCGTGTC TGAAGTTCAT GTGACCACCT CATGAAGCAG AGGAGCGCAC GAGGTGAAGT CATCGACTTC ATCAAGGTGA GCTCGCCGAC TGCTGCCCCG CCCAACGAGA CGGGATCACT CC	CGCCACCATG TCCTGGTCGA GGCGAGGGCG CTGCACCACC TCAGCTACGG CACGACTTCT CATCTTCTTC TCGAGGGCGA AAGGAGGACG GCAACATCCT CACTACCAGC CAACCACTAC AGCGCGATCA CACGGCATGG	GTGAGCAAGG GCTGGACGGC AGGGCGATGC GGCAAGCTGC CGTGCACTGC TCAAGTCCGC AAGGACGACG GCAACATCCT CCGCCACAAC AGAACACCCC CTGAGCACCC CAACCACTAC ACGAGCTGTA	GCGAGGAGCT GACGTGAACG CACCTACGGC CCGTGCCCTG TTCAGCCGCT CATGCCCGAA GCAACTACAA AACCGCATCG GGGGCACAAG CCGACAAGCA ATCGAGGACG CATCGGCGAC AGTCCGCCCT CTGGAGTTCTG CAAGTAAAGC			

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

- 1 -

## HIGH LEVEL EXPRESSION OF PROTEINS

### Field of the Invention

5           The invention concerns genes and methods for expressing eukaryotic and viral proteins at high levels in eukaryotic cells.

### Background of the Invention

          Expression of eukaryotic gene products in  
10 prokaryotes is sometimes limited by the presence of codons that are infrequently used in *E. coli*. Expression of such genes can be enhanced by systematic substitution of the endogenous codons with codons over represented in highly expressed prokaryotic genes (Robinson et al.  
15 1984). It is commonly supposed that rare codons cause pausing of the ribosome, which leads to a failure to complete the nascent polypeptide chain and a uncoupling of transcription and translation. The mRNA 3' end of the stalled ribosome is exposed to cellular ribonucleases,  
20 which decreases the stability of the transcript.

### Summary of the Invention

          The invention features a synthetic gene encoding a protein normally expressed in a mammalian cell or other eukaryotic cell wherein at least one non-preferred or  
25 less preferred codon in the natural gene encoding the protein has been replaced by a preferred codon encoding the same amino acid.

          Preferred codons are: Ala (gcc); Arg (cgc); Asn (aac); Asp (gac) Cys (tgc); Gln (cag); Gly (ggc); His  
30 (cac); Ile (atc); Leu (ctg); Lys (aag); Pro (ccc); Phe (ttc); Ser (agc); Thr (acc); Tyr (tac); and Val (gtg). Less preferred codons are: Gly (ggg); Ile (att); Leu (ctc); Ser (tcc); Val (gtc). All codons which do not fit the description of preferred codons or less preferred  
35 codons are non-preferred codons. In general, the degree

- 2 -

of preference of particular codon is indicated by the prevalence of the codon in highly expressed human genes as indicated in Table 1 under the heading "High." For example, "atc" represents 77% of the Ile codons in highly expressed mammalian genes and is the preferred Ile codon; "att" represents 18% of the Ile codons in highly expressed mammalian genes and is the less preferred Ile codon. The sequence "ata" represents only 5% of the Ile codons in highly expressed human genes as is a non-preferred codon. Replacing a codon with another codon that is more prevalent in highly expressed human genes will generally increase expression of the gene in mammalian cells. Accordingly, the invention includes replacing a less preferred codon with a preferred codon as well as replacing a non-preferred codon with a preferred or less preferred codon.

By "protein normally expressed in a mammalian cell" is meant a protein which is expressed in mammalian under natural conditions. The term includes genes in the mammalian genome such as Factor VIII, Factor IX, interleukins, and other proteins. The term also includes genes which are expressed in a mammalian cell under disease conditions such as oncogenes as well as genes which are encoded by a virus (including a retrovirus) which are expressed in mammalian cells post-infection. By "protein normally expressed in a eukaryotic cell" is meant a protein which is expressed in a eukaryote under natural conditions. The term also includes genes which are expressed in a mammalian cell under disease conditions such as

In preferred embodiments, the synthetic gene is capable of expressing the mammalian or eukaryotic protein at a level which is at least 110%, 150%, 200%, 500%, 1,000%, 5,000% or 10,000% of that expressed by said natural gene in an in vitro mammalian cell culture system

- 3 -

under identical conditions (i.e., same cell type, same culture conditions, same expression vector).

Suitable cell culture systems for measuring expression of the synthetic gene and corresponding  
5 natural gene are described below. Other suitable expression systems employing mammalian cells are well known to those skilled in the art and are described in, for example, the standard molecular biology reference works noted below. Vectors suitable for expressing the  
10 synthetic and natural genes are described below and in the standard reference works described below. By "expression" is meant protein expression. Expression can be measured using an antibody specific for the protein of interest. Such antibodies and measurement techniques are  
15 well known to those skilled in the art. By "natural gene" is meant the gene sequence (including naturally occurring allelic variants) which naturally encodes the protein.

In other preferred embodiments at least 10%, 20%,  
20 30%, 40%, 50%, 60%, 70%, 80%, or 90% of the codons in the natural gene are non-preferred codons.

In a preferred embodiment the protein is a retroviral protein. In a more preferred embodiment the protein is a lentiviral protein. In an even more  
25 preferred embodiment the protein is an HIV protein. In other preferred embodiments the protein is gag, pol, env, gp120, or gp160. In other preferred embodiments the protein is a human protein.

The invention also features a method for preparing  
30 a synthetic gene encoding a protein normally expressed by a mammalian cell or other eukaryotic cell. The method includes identifying non-preferred and less-preferred codons in the natural gene encoding the protein and replacing one or more of the non-preferred and less-

- 4 -

preferred codons with a preferred codon encoding the same amino acid as the replaced codon.

Under some circumstances (e.g., to permit introduction of a restriction site) it may be desirable to replace a non-preferred codon with a less preferred codon rather than a preferred codon.

It is not necessary to replace all less preferred or non-preferred codons with preferred codons. Increased expression can be accomplished even with partial replacement. Under some circumstances it may be desirable to only partially replace non-preferred codons with preferred or less preferred codons in order to obtain an intermediate level of expression.

In other preferred embodiments the invention features vectors (including expression vectors) comprising one or more the synthetic genes.

By "vector" is meant a DNA molecule, derived, e.g., from a plasmid, bacteriophage, or mammalian or insect virus, into which fragments of DNA may be inserted or cloned. A vector will contain one or more unique restriction sites and may be capable of autonomous replication in a defined host or vehicle organism such that the cloned sequence is reproducible. Thus, by "expression vector" is meant any autonomous element capable of directing the synthesis of a protein. Such DNA expression vectors include mammalian plasmids and viruses.

The invention also features synthetic gene fragments which encode a desired portion of the protein. Such synthetic gene fragments are similar to the synthetic genes of the invention except that they encode only a portion of the protein. Such gene fragments preferably encode at least 50, 100, 150, or 500 contiguous amino acids of the protein.

- 5 -

In constructing the synthetic genes of the invention it may be desirable to avoid CpG sequences as these sequences may cause gene silencing.

The codon bias present in the HIV gp120 envelope gene is also present in the gag and pol proteins. Thus, replacement of a portion of the non-preferred and less preferred codons found in these genes with preferred codons should produce a gene capable of higher level expression. A large fraction of the codons in the human genes encoding Factor VIII and Factor IX are non-preferred codons or less preferred codons. Replacement of a portion of these codons with preferred codons should yield genes capable of higher level expression in mammalian cell culture.

The synthetic genes of the invention can be introduced into the cells of a living organism. For example, vectors (viral or non-viral) can be used to introduce a synthetic gene into cells of a living organism for gene therapy.

Conversely, it may be desirable to replace preferred codons in a naturally occurring gene with less-preferred codons as a means of lowering expression.

Standard reference works describing the general principles of recombinant DNA technology include Watson, J.D. et al., Molecular Biology of the Gene, Volumes I and II, the Benjamin/Cummings Publishing Company, Inc., publisher, Menlo Park, CA (1987); Darnell, J.E. et al., Molecular Cell Biology, Scientific American Books, Inc., Publisher, New York, N.Y. (1986); Old, R.W., et al., Principles of Gene Manipulation: An Introduction to Genetic Engineering, 2d edition, University of California Press, publisher, Berkeley, CA (1981); Maniatis, T., et al., Molecular Cloning: A Laboratory Manual, 2nd Ed. Cold Spring Harbor Laboratory, publisher, Cold Spring Harbor, NY (1989); and Current Protocols in Molecular

- 6 -

Biology, Ausubel et al., Wiley Press, New York, NY (1992).

### Detailed Description

#### Description of the Drawings

5           Figure 1 depicts the sequence of the synthetic gp120 and a synthetic gp160 gene in which codons have been replaced by those found in highly expressed human genes.

10           Figure 2 is a schematic drawing of the synthetic gp120 (HIV-1 MN) gene. The shaded portions marked v1 to v5 indicate hypervariable regions. The filled box indicates the CD4 binding site. A limited number of the unique restriction sites are shown: H (Hind3), Nh (Nhe1), P (Pst1), Na (Nae1), M (Mlu1), R (EcoR1), A (Age1) and No (Not1). The chemically synthesized DNA fragments which served as PCR templates are shown below the gp120 sequence, along with the locations of the primers used for their amplification.

15           Figure 3 is a photograph of the results of transient transfection assays used to measure gp120 expression. Gel electrophoresis of immunoprecipitated supernatants of 293T cells transfected with plasmids expressing gp120 encoded by the IIIB isolate of HIV-1 (gp120IIIB), by the MN isolate (gp120mn), by the MN isolate modified by substitution of the endogenous leader peptide with that of the CD5 antigen (gp120mnCD5L), or by the chemically synthesized gene encoding the MN variant with the human CD5Leader (syngp120mn). Supernatants were harvested following a 12 hour labeling period 60 hours post-transfection and immunoprecipitated with CD4:IgG1 fusion protein and protein A sepharose.

20           Figure 4 is a graph depicting the results of ELISA assays used to measure protein levels in supernatants of transiently transfected 293T cells. Supernatants of 293T cells transfected with plasmids expressing gp120 encoded



- 7 -

by the IIIB isolate of HIV-1 (gp120 IIIB), by the MN isolate (gp120mn), by the MN isolate modified by substitution of the endogenous leader peptide with that of CD5 antigen (gp120mn CD5L), or by the chemically  
5 synthesized gene encoding the MN variant with human CDS leader (syngp120mn) were harvested after 4 days and tested in a gp120/CD4 ELISA. The level of gp120 is expressed in ng/ml.

Figure 5, panel A is a photograph of a gel  
10 illustrating the results of a immunoprecipitation assay used to measure expression of the native and synthetic gp120 in the presence of rev in trans and the RRE in cis. In this experiment 293T cells were transiently transfected by calcium phosphate coprecipitation of 10  $\mu$ g  
15 of plasmid expressing: (A) the synthetic gp120MN sequence and RRE in cis, (B) the gp120 portion of HIV-1 IIIB, (C) the gp120 portion of HIV-1 IIIB and RRE in cis, all in the presence or absence of rev expression. The RRE constructs gp120IIIBRRE and syngp120mnRRE were generated  
20 using an EagI/HpaI RRE fragment cloned by PCR from a HIV-1 HXB2 proviral clone. Each gp120 expression plasmid was cotransfected with 10  $\mu$ g of either pCMVrev or CDM7 plasmid DNA. Supernatants were harvested 60 hours post transfection, immunoprecipitated with CD4:IgG fusion  
25 protein and protein A agarose, and run on a 7% reducing SDS-PAGE. The gel exposure time was extended to allow the induction of gp120IIIBrre by rev to be demonstrated.

Figure 5, panel B is a shorter exposure of a similar experiment in which syngp120mnrrre was  
30 cotransfected with or without pCMVrev. Figure 5, panel C is a schematic diagram of the constructs used in panel A.

Figure 6 is a comparison of the sequence of the wild-type rat THY-1 gene (wt) and a synthetic rat THY-1 gene (env) constructed by chemical synthesis and having  
35 the most prevalent codons found in the HIV-1 env gene.

- 8 -

Figure 7 is a schematic diagram of the synthetic ratTHY-1 gene. The solid black box denotes the signal peptide. The shaded box denotes the sequences in the precursor which direct the attachment of a phosphatidyl-  
inositol glycan anchor. Unique restriction sites used for assembly of the THY-1 constructs are marked H (Hind3), M (Mlu1), S (Sac1) and No (Not1). The position of the synthetic oligonucleotides employed in the construction are shown at the bottom of the figure.

Figure 8 is a graph depicting the results of flow cytometry analysis. In this experiment 293T cells transiently transfected with either wild-type rat THY-1 (dark line), ratTHY-1 with envelope codons (light line) or vector only (dotted line). 293T cells were transfected with the different expression plasmids by calcium phosphate coprecipitation and stained with anti-ratTHY-1 monoclonal antibody OX7 followed by a polyclonal FITC- conjugated anti-mouse IgG antibody 3 days after transfection.

Figure 9, panel A is a photograph of a gel illustrating the results of immunoprecipitation analysis of supernatants of human 293T cells transfected with either syngp120mn (A) or a construct syngp120mn.rTHY-1env which has the rTHY-1env gene in the 3' untranslated region of the syngp120mn gene (B). The syngp120mn.rTHY-1env construct was generated by inserting a Not1 adapter into the blunted Hind3 site of the rTHY-1env plasmid. Subsequently, a 0.5 kb Not1 fragment containing the rTHY-1env gene was cloned into the Not1 site of the syngp120mn plasmid and tested for correct orientation. Supernatants of <sup>35</sup>S labeled cells were harvested 72 hours post transfection, precipitated with CD4:IgG fusion protein and protein A agarose, and run on a 7% reducing SDS-PAGE.

- 9 -

Figure 9, panel B is a schematic diagram of the constructs used in the experiment depicted in panel A of FIG. 9.

Figure 10, panel A is a photograph of COS cells  
5 transfected with vector only showing no GFP fluorescence.  
Figure 10, panel B is a photograph of COS cells  
transfected with a CDM7 expression plasmid encoding  
native GFP engineered to include a consensus  
translational initiation sequence. Figure 10, panel C is  
10 a photograph of COS cells transfected with an expression  
plasmid having the same flanking sequences and initiation  
consensus as in FIG. 10, panel B, but bearing a codon  
optimized gene sequence. Figure 10, panel D is a  
photograph of COS cells transfected with an expression  
15 plasmid as in FIG. 10, panel C, but bearing a Thr at  
residue 65 in place of Ser.

Description of the Preferred Embodiments  
Construction of a Synthetic gp120 Gene Having Codons  
Found in Highly Expressed Human Genes

20 A codon frequency table for the envelope precursor  
of the LAV subtype of HIV-1 was generated using software  
developed by the University of Wisconsin Genetics  
Computer Group. The results of that tabulation are  
contrasted in Table 1 with the pattern of codon usage by  
25 a collection of highly expressed human genes. For any  
amino acid encoded by degenerate codons, the most favored  
codon of the highly expressed genes is different from the  
most favored codon of the HIV envelope precursor.  
Moreover a simple rule describes the pattern of favored  
30 envelope codons wherever it applies: preferred codons  
maximize the number of  
adenine residues in the viral RNA. In all cases but one  
this means that the codon in which the third position is  
A is the most frequently used. In the special case of  
35 serine, three codons equally contribute one A residue to

- 10 -

the mRNA; together these three comprise 85% of the serine codons actually used in envelope transcripts. A particularly striking example of the A bias is found in the codon choice for arginine, in which the AGA triplet  
5 comprises 88% of the arginine codons. In addition to the preponderance of A residues, a marked preference is seen for uridine among degenerate codons whose third residue must be a pyrimidine. Finally, the inconsistencies among the less frequently used variants can be accounted for by  
10 the observation that the dinucleotide CpG is under represented; thus the third position is less likely to be G whenever the second position is C, as in the codons for alanine, proline, serine and threonine; and the CGX triplets for arginine are hardly used at all.

- 11 -

**TABLE 1:** Codon Frequency in the HIV-1 IIIb env gene and in highly expressed human genes.

		High	Env			High	Env
<u>Ala</u>				<u>Cys</u>			
GC	C	53	27	TG	C	68	16
	T	17	18		T	32	84
	A	13	50	<u>Gln</u>			
	G	17	5	CA	A	12	55
<u>Arg</u>					G	88	45
CG	C	37	0	<u>Glu</u>			
	T	7	4	GA	A	25	67
	A	6	0		G	75	33
	G	21	0	<u>Gly</u>			
AG	A	10	88	GG	C	50	6
	G	18	8		T	12	13
<u>Asn</u>					A	14	53
AA	C	78	30		G	24	28
	T	22	70	<u>His</u>			
<u>Asp</u>				CA	C	79	25
GA	C	75	33		T	21	75
	T	25	67	<u>Ile</u>			
<u>Leu</u>				AT	C	77	25
CT	C	26	10		T	18	31
	T	5	7	A	5	44	
	A	3	17	<u>Ser</u>			
	G	58	17	TC	C	28	8
TT	A	2	30		T	13	8
	G	6	20	A	5	22	
<u>Lys</u>				G	9	0	
AA	A	18	68	AG	C	34	22
	G	82	32	T	10	41	
<u>Pro</u>				<u>Thr</u>			
CC	C	48	27	AC	C	57	20
	T	19	14		T	14	22
	A	16	55	A	14	51	
	G	17	5	G	15	7	
<u>Phe</u>				<u>Tyr</u>			
TT	C	80	26	TA	C	74	8
	T	20	74		T	26	92
				<u>Val</u>			
				GT	C	25	12
					T	7	9
					A	5	62
					G	64	18

Codon frequency was calculated using the GCG program established the University of Wisconsin Genetics Computer Group. Numbers represent the percentage of cases in which the particular codon is used. Codon usage frequencies of envelope genes of other HIV-1 virus isolates are comparable and show a similar bias.

- 12 -

In order to produce a gp120 gene capable of high level expression in mammalian cells, a synthetic gene encoding the gp120 segment of HIV-1 was constructed (syngp120mn), based on the sequence of the most common North American subtype, HIV-1 MN (Shaw et al., Science 226:1165, 1984; Gallo et al., Nature 321:119, 1986). In this synthetic gp120 gene nearly all of the native codons have been systematically replaced with codons most frequently used in highly expressed human genes (FIG. 1).

This synthetic gene was assembled from chemically synthesized oligonucleotides of 150 to 200 bases in length. If oligonucleotides exceeding 120 to 150 bases are chemically synthesized, the percentage of full-length product can be low, and the vast excess of material consists of shorter oligonucleotides. Since these shorter fragments inhibit cloning and PCR procedures, it can be very difficult to use oligonucleotides exceeding a certain length. In order to use crude synthesis material without prior purification, single-stranded oligonucleotide pools were PCR amplified before cloning. PCR products were purified in agarose gels and used as templates in the next PCR step. Two adjacent fragments could be co-amplified because of overlapping sequences at the end of either fragment. These fragments, which were between 350 and 400 bp in size, were subcloned into a pCDM7-derived plasmid containing the leader sequence of the CD5 surface molecule followed by a Nhe1/Pst1/Mlu1/EcoR1/BamH1 polylinker. Each of the restriction enzymes in this polylinker represents a site that is present at either the 5' or 3' end of the PCR-generated fragments. Thus, by sequential subcloning of each of the 4 long fragments, the whole gp120 gene was assembled. For each fragment 3 to 6 different clones were subcloned and sequenced prior to assembly. A schematic drawing of the method used to construct the synthetic

- 13 -

gp120 is shown in FIG. 2. The sequence of the synthetic gp120 gene (and a synthetic gp160 gene created using the same approach) is presented in FIG. 1.

The mutation rate was considerable. The most  
5 commonly found mutations were short (1 nucleotide) and long (up to 30 nucleotides) deletions. In some cases it was necessary to exchange parts with either synthetic adapters or pieces from other subclones without mutation in that particular region. Some deviations from strict  
10 adherence to optimized codon usage were made to accommodate the introduction of restriction sites into the resulting gene to facilitate the replacement of various segments (FIG. 2). These unique restriction sites were introduced into the gene at approximately 100 bp  
15 intervals. The native HIV leader sequence was exchanged with the highly efficient leader peptide of the human CD5 antigen to facilitate secretion (Aruffo et al., Cell.61:1303, 1990) The plasmid used for construction is a derivative of the mammalian expression vector pCDM7  
20 transcribing the inserted gene under the control of a strong human CMV immediate early promoter.

To compare the wild-type and synthetic gp120 coding sequences, the synthetic gp120 coding sequence was inserted into a mammalian expression vector and tested in  
25 transient transfection assays. Several different native gp120 genes were used as controls to exclude variations in expression levels between different virus isolates and artifacts induced by distinct leader sequences. The gp120 HIV IIIB construct used as control was generated by  
30 PCR using a Sall/Xho1 HIV-1 HXB2 envelope fragment as template. To exclude PCR induced mutations, a Kpn1/Ear1 fragment containing approximately 1.2 kb of the gene was exchanged with the respective sequence from the proviral clone. The wild-type gp120mn constructs used as controls  
35 were cloned by PCR from HIV-1 MN infected C8166 cells

- 14 -

(AIDS Repository, Rockville, MD) and expressed gp120 either with a native envelope or a CD5 leader sequence. Since proviral clones were not available in this case, two clones of each construct were tested to avoid PCR artifacts. To determine the amount of secreted gp120 semi-quantitatively supernatants of 293T cells transiently transfected by calcium phosphate co-precipitation were immunoprecipitated with soluble CD4:immunoglobulin fusion protein and protein A sepharose.

The results of this analysis (FIG. 3) show that the synthetic gene product is expressed at a very high level compared to that of the native gp120 controls. The molecular weight of the synthetic gp120 gene was comparable to control proteins (FIG. 3) and appeared to be in the range of 100 to 110 kd. The slightly faster migration can be explained by the fact that in some tumor cell lines like 293T glycosylation is either not complete or altered to some extent.

To compare expression more accurately gp120 protein levels were quantitated using a gp120 ELISA with CD4 in the demobilized phase. This analysis shows (FIG. 4) that ELISA data were comparable to the immunoprecipitation data, with a gp120 concentration of approximately 125 ng/ml for the synthetic gp120 gene, and less than the background cutoff (5 ng/ml) for all the native gp120 genes. Thus, expression of the synthetic gp120 gene appears to be at least one order of magnitude higher than wild-type gp120 genes. In the experiment shown the increase was at least 25 fold.

#### The Role of rev in gp120 Expression

Since rev appears to exert its effect at several steps in the expression of a viral transcript, the possible role of non-translational effects in the improved expression of the synthetic gp120 gene was



- 15 -

tested. First, to rule out the possibility that negative signals elements conferring either increased mRNA degradation or nucleic retention were eliminated by changing the nucleotide sequence, cytoplasmic mRNA levels were tested. Cytoplasmic RNA was prepared by NP40 lysis of transiently transfected 293T cells and subsequent elimination of the nuclei by centrifugation. Cytoplasmic RNA was subsequently prepared from lysates by multiple phenol extractions and precipitation, spotted on nitrocellulose using a slot blot apparatus, and finally hybridized with an envelope-specific probe.

Briefly, cytoplasmic mRNA 293 cells transfected with CDM8, gp120 IIIB, or syngp120 was isolated 36 hours post transfection. Cytoplasmic RNA of Hela cells infected with wild-type vaccinia virus or recombinant virus expressing gp120 IIIB or the synthetic gp120 gene was under the control of the 7.5 promoter was isolated 16 hours post infection. Equal amounts were spotted on nitrocellulose using a slot blot device and hybridized with randomly labeled 1.5 kb gp120IIIB and syngp120 fragments or human beta-actin. RNA expression levels were quantitated by scanning the hybridized membranes with a phosphorimager. The procedures used are described in greater detail below.

This experiment demonstrated that there was no significant difference in the mRNA levels of cells transfected with either the native or synthetic gp120 gene. In fact, in some experiments cytoplasmic mRNA level of the synthetic gp120 gene was even lower than that of the native gp120 gene.

These data were confirmed by measuring expression from recombinant vaccinia viruses. Human 293 cells or Hela cells were infected with vaccinia virus expressing wild-type gp120 IIIB or syngp120mn at a multiplicity of infection of at least 10. Supernatants were harvested

- 16 -

hours post infection and immunoprecipitated with CD4:immunoglobulin fusion protein and protein A sepharose. The procedures used in this experiment are described in greater detail below.

5           This experiment showed that the increased expression of the synthetic gene was still observed when the endogenous gene product and the synthetic gene product were expressed from vaccinia virus recombinants under the control of the strong mixed early and late 7.5k  
10 promoter. Because vaccinia virus mRNAs are transcribed and translated in the cytoplasm, increased expression of the synthetic envelope gene in this experiment cannot be attributed to improved export from the nucleus. This  
15 experiment was repeated in two additional human cell types, the kidney cancer cell line 293 and HeLa cells. As with transfected 293T cells, mRNA levels were similar in 293 cells infected with either recombinant vaccinia virus.

#### Codon Usage in Lentivirus

20           Because it appears that codon usage has a significant impact on expression in mammalian cells, the codon frequency in the envelope genes of other  
retroviruses was examined. This study found no clear pattern of codon preference between retroviruses in  
25 general. However, if viruses from the lentivirus genus, to which HIV-1 belongs to, were analyzed separately, codon usage bias almost identical to that of HIV-1 was found. A codon frequency table from the envelope  
glycoproteins of a variety of (predominantly type C)  
30 retroviruses excluding the lentiviruses was prepared, and compared a codon frequency table created from the envelope sequences of four lentiviruses not closely related to HIV-1 (caprine arthritis encephalitis virus, equine infectious anemia virus, feline immunodeficiency  
35 virus, and visna virus) (Table 2). The codon usage

- 17 -

pattern for lentiviruses is strikingly similar to that of HIV-1, in all cases but one, the preferred codon for HIV-1 is the same as the preferred codon for the other lentiviruses. The exception is proline, which is encoded  
5 by CCT in 41% of non-HIV lentiviral envelope residues, and by CCA in 40% of residues, a situation which clearly also reflects a significant preference for the triplet ending in A. The pattern of codon usage by the non-lentiviral envelope proteins does not show a similar  
10 predominance of A residues, and is also not as skewed toward third position C and G residues as is the codon usage for the highly expressed human genes. In general non-lentiviral retroviruses appear to exploit the different codons more equally, a pattern they share with  
15 less highly expressed human genes.

- 18 -

**TABLE 2:** Codon frequency in the envelope gene of lentiviruses (lenti) and non-lentiviral retroviruses (other).

Other Lenti				Other Lenti			
<u>Ala</u>				<u>Cys</u>			
GC	C	45	13	TG	C	53	21
	T	26	37		T	47	79
	A	20	46				
	G	9	3	<u>Gln</u>			
<u>Arg</u>				CA	A	52	69
CG	C	14	2		G	48	31
	T	6	3	<u>Glu</u>			
	A	16	5	GA	A	57	68
	G	17	3		G	43	32
AG	A	31	51	<u>Gly</u>			
	G	15	26	GG	C	21	8
<u>Asn</u>					T	13	9
AA	C	49	31		A	37	56
	T	51	69		G	29	26
<u>Asp</u>				<u>His</u>			
GA	C	55	33	CA	C	51	38
	T	51	69		T	49	62
<u>Leu</u>				<u>Ile</u>			
CT	C	22	8	AT	C	38	16
	T	14	9		T	31	22
	A	21	16		A	31	61
	G	19	11	<u>Ser</u>			
TT	A	15	41	TC	C	38	10
	G	10	16		T	17	16
<u>Lys</u>					A	18	24
AA	A	60	63		G	6	5
	G	40	37	AG	C	13	20
<u>Pro</u>					T	7	25
CC	C	42	14	<u>Thr</u>			
	T	30	41	AC	C	44	18
	A	20	40		T	27	20
	G	7	5		A	19	55
<u>Phe</u>					G	10	8
TT	C	52	25	<u>Tyr</u>			
	T	48	75	TA	C	48	28
					T	52	72
				<u>Val</u>			
				GT	C	36	9
					T	17	10
					A	22	54
					G	25	27

Codon frequency was calculated using the GCG program established by the University of Wisconsin Genetics Computer Group. Numbers represent the percentage in which a particular codon is used. Codon usage of non-lentiviral retroviruses was compiled from the envelope precursor sequences of bovine leukemia virus, feline leukemia virus, human T-cell leukemia virus type I, human T-cell lymphotropic virus type II, the mink cell focus-forming isolate of murine leukemia virus (MuLV), the Rauscher spleen focus-forming isolate, the 10A1 isolate, the 4070A amphotropic isolate and the myeloproliferative leukemia

- 19 -

virus isolate, and from rat leukemia virus, simian sarcoma virus, simian T-cell leukemia virus, leukemogenic retrovirus T1223/B and gibbon ape leukemia virus. The codon frequency tables for the non-HIV, non-SIV lentiviruses were compiled from the envelope precursor sequences for caprine arthritis encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus.

---

In addition to the prevalence of A containing codons, lentiviral codons adhere to the HIV pattern of strong CpG under representation, so that the third position for alanine, proline, serine and threonine  
5 triplets is rarely G. The retroviral envelope triplets show a similar, but less pronounced, under representation of CpG. The most obvious difference between lentiviruses and other retroviruses with respect to CpG prevalence lies in the usage of the CGX variant of arginine  
10 triplets, which is reasonably frequently represented among the retroviral envelope coding sequences, but is almost never present among the comparable lentivirus sequences.

15 Differences in rev Dependence Between Native and Synthetic gp120

To examine whether regulation by rev is connected to HIV-1 codon usage, the influence of rev on the expression of both native and synthetic gene was investigated. Since regulation by rev requires the rev-  
20 binding site RRE in cis, constructs were made in which this binding site was cloned into the 3' untranslated region of both the native and the synthetic gene. These plasmids were co-transfected with rev or a control plasmid in trans into 293T cells, and gp120 expression  
25 levels in supernatants were measured semiquantitatively by immunoprecipitation. The procedures used in this experiment are described in greater detail below.

As shown in FIG. 5, panel A and FIG. 5, panel B, rev up regulates the native gp120 gene, but has no effect  
30 on the expression of the synthetic gp120 gene. Thus, the

- 20 -

action of rev is not apparent on a substrate which lacks the coding sequence of endogenous viral envelope sequences.

5 Expression of a synthetic rat THY-1 gene with HIV envelope codons

The above-described experiment suggest that in fact "envelope sequences" have to be present for rev regulation. In order to test this hypothesis, a  
10 synthetic version of the gene encoding the small, typically highly expressed cell surface protein, rat THY-1 antigen, was prepared. The synthetic version of the rat THY-1 gene was designed to have a codon usage like that of HIV gp120. In designing this synthetic gene  
15 AUUUA sequences, which are associated with mRNA instability, were avoided. In addition, two restriction sites were introduced to simplify manipulation of the resulting gene (FIG 6). This synthetic gene with the HIV envelope codon usage (rTHY-1env) was generated using  
20 three 150 to 170 mer oligonucleotides (FIG. 7). In contrast to the syngp120mn gene, PCR products were directly cloned and assembled in pUC12, and subsequently cloned into pCDM7.

Expression levels of native rTHY-1 and rTHY-1 with  
25 the HIV envelope codons were quantitated by immunofluorescence of transiently transfected 293T cells. FIG 8 shows that the expression of the native THY-1 gene is almost two orders of magnitude above the background level of the control transfected cells (pCDM7). In  
30 contrast, expression of the synthetic rat THY-1 is substantially lower than that of the native gene (shown by the shift to of the peak towards a lower channel number).

To prove that no negative sequence elements  
35 promoting mRNA degradation were inadvertently introduced, a construct was generated in which the rTHY-1env gene was

- 21 -

cloned at the 3' end of the synthetic gp120 gene (FIG. 9, panel B). In this experiment 293T cells were transfected with either the syngp120mn gene or the syngp120/rat THY-1 env fusion gene (syngp120mn.rTHY-1env). Expression was  
5 measured by immunoprecipitation with CD4:IgG fusion protein and protein A agarose. The procedures used in this experiment are described in greater detail below.

Since the synthetic gp120 gene has an UAG stop codon, rTHY-1env is not translated from this transcript.  
10 If negative elements conferring enhanced degradation were present in the sequence, gp120 protein levels expressed from this construct should be decreased in comparison to the syngp120mn construct without rTHY-1env. FIG. 9, panel A, shows that the expression of both constructs is  
15 similar, indicating that the low expression must be linked to translation.

Rev-dependent expression of synthetic rat THY-1 gene with envelope codons

To explore whether rev is able to regulate  
20 expression of a rat THY-1 gene having env codons, a construct was made with a rev-binding site in the 3' end of the rTHY1env open reading frame. To measure rev-responsiveness of the a rat THY-1env construct having a 3' RRE, human 293T cells were cotransfected  
25 ratTHY-1envrre and either CDM7 or pCMVrev. At 60 hours post transfection cells were detached with 1 mM EDTA in PBS and stained with the OX-7 anti rTHY-1 mouse monoclonal antibody and a secondary FITC-conjugated antibody. Fluorescence intensity was measured using a  
30 EPICS XL cytofluorometer. These procedures are described in greater detail below.

In repeated experiments, a slight increase of rTHY-1env expression was detected if rev was cotransfected with the rTHY-1env gene. To further  
35 increase the sensitivity of the assay system a construct

- 22 -

expressing a secreted version of rTHY-1env was generated. This construct should produce more reliable data because the accumulated amount of secreted protein in the supernatant reflects the result of protein production over an extended period, in contrast to surface expressed protein, which appears to more closely reflect the current production rate. A gene capable of expressing a secreted form was prepared by PCR using forward and reverse primers annealing 3' of the endogenous leader sequence and 5' of the sequence motif required for phosphatidylinositol glycan anchorage respectively. The PCR product was cloned into a plasmid which already contained a CD5 leader sequence, thus generating a construct in which the membrane anchor has been deleted and the leader sequence exchanged by a heterologous (and probably more efficient) leader peptide.

The rev-responsiveness of the secreted form ratTHY-1env was measured by immunoprecipitation of supernatants of human 293T cells cotransfected with a plasmid expressing a secreted form of ratTHY-1env and the RRE sequence in cis (rTHY-1envPI-rre) and either CDM7 or pCMVrev. The rTHY-1envPI-RRE construct was made by PCR using the oligonucleotide: cgcggggctagcgcaaagagtaataagtttaac (SEQ ID NO:38) as a forward primer, the oligonucleotide: cgcggatcccttgatattttgtactaata (SEQ ID NO:39) as reverse primer, and the synthetic rTHY-1env construct as template. After digestion with NheI and NotI the PCR fragment was cloned into a plasmid containing CD5 leader and RRE sequences. Supernatants of <sup>35</sup>S labeled cells were harvested 72 hours post transfection, precipitated with a mouse monoclonal antibody OX7 against rTHY-1 and anti mouse IgG sepharose, and run on a 12% reducing SDS-PAGE.



- 23 -

In this experiment the induction of rTHY-1env by rev was much more prominent and clear-cut than in the above-described experiment and strongly suggests that rev is able to translationally regulate transcripts that are suppressed by low-usage codons.

Rev-independent expression of a rTHY-1env:immunoglobulin fusion protein

To test whether low-usage codons must be present throughout the whole coding sequence or whether a short region is sufficient to confer rev-responsiveness, a rTHY-1env:immunoglobulin fusion protein was generated. In this construct the rTHY-1env gene (without the sequence motif responsible for phosphatidylinositol glycan anchorage) is linked to the human IgG1 hinge, CH2 and CH3 domains. This construct was generated by anchor PCR using primers with NheI and BamHI restriction sites and rTHY-1env as template. The PCR fragment was cloned into a plasmid containing the leader sequence of the CD5 surface molecule and the hinge, CH2 and CH3 parts of human IgG1 immunoglobulin. A Hind3/EagI fragment containing the rTHY-1env<sup>egl</sup> insert was subsequently cloned into a pCDM7-derived plasmid with the RRE sequence.

To measure the response of the rTHY-1env/immunoglobulin fusion gene (rTHY-1env<sup>eglrr</sup>) to rev human 293T cells cotransfected with rTHY-1env<sup>eglrr</sup> and either pCDM7 or pCMVrev. The rTHY-1env<sup>eglrr</sup> construct was made by anchor PCR using forward and reverse primers with NheI and BamHI restriction sites respectively. The PCR fragment was cloned into a plasmid containing a CD5 leader and human IgG1 hinge, CH2 and CH3 domains. Supernatants of <sup>35</sup>S labeled cells were harvested 72 hours post transfection, precipitated with a mouse monoclonal antibody OX7 against rTHY-1 and anti mouse IgG sepharose, and run on a 12% reducing SDS-PAGE. The procedures used are described in greater detail below.

- 24 -

As with the product of the rTHY-1envPI- gene, this rTHY-1env/immunoglobulin fusion protein is secreted into the supernatant. Thus, this gene should be responsive to rev-induction. However, in contrast to rTHY-1envPI-,  
5 cotransfection of rev in trans induced no or only a negligible increase of rTHY-1envveg1 expression.

The expression of rTHY-1:immunoglobulin fusion protein with native rTHY-1 or HIV envelope codons was measured by immunoprecipitation. Briefly, human 293T  
10 cells transfected with either rTHY-1envveg1 (env codons) or rTHY-1wtegl1 (native codons). The rTHY-1wtegl1 construct was generated in manner similar to that used for the rTHY-1envveg1 construct, with the exception that a plasmid containing the native rTHY-1 gene was used as  
15 template. Supernatants of <sup>35</sup>S labeled cells were harvested 72 hours post transfection, precipitated with a mouse monoclonal antibody OX7 against rTHY-1 and anti mouse IgG sepharose, and run on a 12% reducing SDS-PAGE. THE procedures used in this experiment are described in  
20 greater detail below.

Expression levels of rTHY-1envveg1 were decreased in comparison to a similar construct with wild-type rTHY-1 as the fusion partner, but were still considerably higher than rTHY-1env. Accordingly, both parts of the  
25 fusion protein influenced expression levels. The addition of rTHY-1env did not restrict expression to an equal level as seen for rTHY-1env alone. Thus, regulation by rev appears to be ineffective if protein expression is not almost completely suppressed.

### 30 Codon preference in HIV-1 envelope genes

Direct comparison between codon usage frequency of HIV envelope and highly expressed human genes reveals a striking difference for all twenty amino acids. One simple measure of the statistical significance of this  
35 codon preference is the finding that among the nine amino

- 25 -

acids with two fold codon degeneracy, the favored third residue is A or U in all nine. The probability that all nine of two equiprobable choices will be the same is approximately 0.004, and hence by any conventional  
5 measure the third residue choice cannot be considered random. Further evidence of a skewed codon preference is found among the more degenerate codons, where a strong selection for triplets bearing adenine can be seen. This contrasts with the pattern for highly expressed genes,  
10 which favor codons bearing C, or less commonly G, in the third position of codons with three or more fold degeneracy.

The systematic exchange of native codons with codons of highly expressed human genes dramatically  
15 increased expression of gp120. A quantitative analysis by ELISA showed that expression of the synthetic gene was at least 25 fold higher in comparison to native gp120 after transient transfection into human 293 cells. The concentration levels in the ELISA experiment shown were  
20 rather low. Since an ELISA was used for quantification which is based on gp120 binding to CD4, only native, non-denatured material was detected. This may explain the apparent low expression. Measurement of cytoplasmic mRNA levels demonstrated that the difference in protein  
25 expression is due to translational differences and not mRNA stability.

Retroviruses in general do not show a similar preference towards A and T as found for HIV. But if this family was divided into two subgroups, lentiviruses and  
30 non-lentiviral retroviruses, a similar preference to A and, less frequently, T, was detected at the third codon position for lentiviruses. Thus, the availing evidence suggests that lentiviruses retain a characteristic pattern of envelope codons not because of an inherent  
35 advantage to the reverse transcription or replication of

- 26 -

such residues, but rather for some reason peculiar to the physiology of that class of viruses. The major difference between lentiviruses and non-complex retroviruses are additional regulatory and non-essentially accessory genes in lentiviruses, as already mentioned. Thus, one simple explanation for the restriction of envelope expression might be that an important regulatory mechanism of one of these additional molecules is based on it. In fact, it is known that one of these proteins, rev, which most likely has homologues in all lentiviruses. Thus codon usage in viral mRNA is used to create a class of transcripts which is susceptible to the stimulatory action of rev. This hypothesis was proved using a similar strategy as above, but this time codon usage was changed into the inverse direction. Codon usage of a highly expressed cellular gene was substituted with the most frequently used codons in the HIV envelope. As assumed, expression levels were considerably lower in comparison to the native molecule, almost two orders of magnitude when analyzed by immunofluorescence of the surface expressed molecule (see 4.7). If rev was coexpressed in trans and a RRE element was present in cis only a slight induction was found for the surface molecule. However, if THY-1 was expressed as a secreted molecule, the induction by rev was much more prominent, supporting the above hypothesis. This can probably be explained by accumulation of secreted protein in the supernatant, which considerably amplifies the rev effect. If rev only induces a minor increase for surface molecules in general, induction of HIV envelope by rev cannot have the purpose of an increased surface abundance, but rather of an increased intracellular gp160 level. It is completely unclear at the moment why this should be the case.

- 27 -

To test whether small subtotal elements of a gene are sufficient to restrict expression and render it rev-dependent rTHY1env:immunoglobulin fusion proteins were generated, in which only about one third of the total  
5 gene had the envelope codon usage. Expression levels of this construct were on an intermediate level, indicating that the rTHY-1env negative sequence element is not dominant over the immunoglobulin part. This fusion protein was not or only slightly rev-responsive,  
10 indicating that only genes almost completely suppressed can be rev-responsive.

Another characteristic feature that was found in the codon frequency tables is a striking under representation of CpG triplets. In a comparative study  
15 of codon usage in E. coli, yeast, drosophila and primates it was shown that in a high number of analyzed primate genes the 8 least used codons contain all codons with the CpG dinucleotide sequence. Avoidance of codons containing this dinucleotide motif was also found in the  
20 sequence of other retroviruses. It seems plausible that the reason for under representation of CpG-bearing triplets has something to do with avoidance of gene silencing by methylation of CpG cytosines. The expected number of CpG dinucleotides for HIV as a whole is about  
25 one fifth that expected on the basis of the base composition. This might indicate that the possibility of high expression is restored, and that the gene in fact has to be highly expressed at some point during viral pathogenesis.

30 The results presented herein clearly indicate that codon preference has a severe effect on protein levels, and suggest that translational elongation is controlling mammalian gene expression. However, other factors may play a role. First, abundance of not maximally loaded  
35 mRNA's in eukaryotic cells indicates that initiation is

- 28 -

rate limiting for translation in at least some cases, since otherwise all transcripts would be completely covered by ribosomes. Furthermore, if ribosome stalling and subsequent mRNA degradation were the mechanism, suppression by rare codons could most likely not be reversed by any regulatory mechanism like the one presented herein. One possible explanation for the influence of both initiation and elongation on translational activity is that the rate of initiation, or access to ribosomes, is controlled in part by cues distributed throughout the RNA, such that the lentiviral codons predispose the RNA to accumulate in a pool of poorly initiated RNAs. However, this limitation need not be kinetic; for example, the choice of codons could influence the probability that a given translation product, once initiated, is properly completed. Under this mechanism, abundance of less favored codons would incur a significant cumulative probability of failure to complete the nascent polypeptide chain. The sequestered RNA would then be lent an improved rate of initiation by the action of rev. Since adenine residues are abundant in rev-responsive transcripts, it could be that RNA adenine methylation mediates this translational suppression.

## 25 Detailed Procedures

The following procedures were used in the above-described experiments.

### Sequence Analysis

Sequence analyses employed the software developed by the University of Wisconsin Computer Group.

### Plasmid constructions

Plasmid constructions employed the following methods. Vectors and insert DNA was digested at a concentration of 0.5  $\mu\text{g}/10\ \mu\text{l}$  in the appropriate restriction buffer for 1 - 4 hours (total reaction volume

- 29 -

approximately 30  $\mu$ l). Digested vector was treated with 10% (v/v) of 1  $\mu$ g/ml calf intestine alkaline phosphatase for 30 min prior to gel electrophoresis. Both vector and insert digests (5 to 10  $\mu$ l each) were run on a 1.5% low melting agarose gel with TAE buffer. Gel slices containing bands of interest were transferred into a 1.5 ml reaction tube, melted at 65°C and directly added to the ligation without removal of the agarose. Ligations were typically done in a total volume of 25  $\mu$ l in 1x Low Buffer 1x Ligation Additions with 200-400 U of ligase, 1  $\mu$ l of vector, and 4  $\mu$ l of insert. When necessary, 5' overhanging ends were filled by adding 1/10 volume of 250  $\mu$ M dNTPs and 2-5 U of Klenow polymerase to heat inactivated or phenol extracted digests and incubating for approximately 20 min at room temperature. When necessary, 3' overhanging ends were filled by adding 1/10 volume of 2.5 mM dNTPs and 5-10 U of T4 DNA polymerase to heat inactivated or phenol extracted digests, followed by incubation at 37°C for 30 min. The following buffers were used in these reactions: 10x Low buffer (60 mM Tris HCl, pH 7.5, 60 mM MgCl<sub>2</sub>, 50 mM NaCl, 4 mg/ml BSA, 70 mM  $\beta$ -mercaptoethanol, 0.02% NaN<sub>3</sub>); 10x Medium buffer (60 mM Tris HCl, pH 7.5, 60 mM MgCl<sub>2</sub>, 50 mM NaCl, 4 mg/ml BSA, 70 mM  $\beta$ -mercaptoethanol, 0.02% NaN<sub>3</sub>); 10x High buffer (60 mM Tris HCl, pH 7.5, 60 mM MgCl<sub>2</sub>, 50 mM NaCl, 4 mg/ml BSA, 70 mM  $\beta$ -mercaptoethanol, 0.02% NaN<sub>3</sub>); 10x Ligation additions (1 mM ATP, 20 mM DTT, 1 mg/ml BSA, 10 mM spermidine); 50x TAE (2 M Tris acetate, 50 mM EDTA).

#### Oligonucleotide synthesis and purification

Oligonucleotides were produced on a Milligen 8750 synthesizer (Millipore). The columns were eluted with 1 ml of 30% ammonium hydroxide, and the eluted oligonucleotides were deblocked at 55°C for 6 to 12 hours. After deblocking, 150  $\mu$ l of oligonucleotide were precipitated with 10x volume of unsaturated n-butanol in

- 30 -

1.5 ml reaction tubes, followed by centrifugation at 15,000 rpm in a microfuge. The pellet was washed with 70% ethanol and resuspended in 50  $\mu$ l of H<sub>2</sub>O. The concentration was determined by measuring the optical density at 260 nm in a dilution of 1:333 (1 OD<sub>260</sub> = 30  $\mu$ g/ml).

The following oligonucleotides were used for construction of the synthetic gp120 gene (all sequences shown in this text are in 5' to 3' direction).

10        oligo 1 forward (Nhe1): cgc ggg cta gcc acc gag aag ctg (SEQ ID NO:1).

         oligo 1: acc gag aag ctg tgg gtg acc gtg tac tac ggc gtg ccc gtg tgg aag ag ag gcc acc acc acc ctg ttc tgc gcc agc gac gcc aag gcg tac gac acc gag gtg cac aac gtg  
15 tgg gcc acc cag gcg tgc gtg ccc acc gac ccc aac ccc cag gag gtg gag ctc gtg aac gtg acc gag aac ttc aac at (SEQ ID NO:2).

         oligo 1 reverse: cca cca tgt tgt tct tcc aca tgt tga agt tct c (SEQ ID NO:3).

20        oligo 2 forward: gac cga gaa ctt caa cat gtg gaa gaa caa cat (SEQ ID NO:4)

         oligo 2: tgg aag aac aac atg gtg gag cag atg cat gag gac atc atc agc ctg tgg gac cag agc ctg aag ccc tgc gtg aag ctg acc cc ctg tgc gtg acc tg aac tgc acc gac ctg  
25 agg aac acc acc aac acc aac ac agc acc gcc aac aac aac agc aac agc gag ggc acc atc aag ggc ggc gag atg (SEQ ID NO:5).

         oligo 2 reverse (Pst1): gtt gaa gct gca gtt ctt cat ctc gcc gcc ctt (SEQ ID NO:6).

30        oligo 3 forward (Pst1): gaa gaa ctg cag ctt caa cat cac cac cag c (SEQ ID NO:7).

         oligo 3: aac atc acc acc agc atc cgc gac aag atg cag aag gag tac gcc ctg ctg tac aag ctg gat atc gtg agc atc gac aac gac agc acc agc tac cgc ctg atc tcc tgc aac  
35 acc agc gtg atc acc cag gcc tgc ccc aag atc agc ttc gag



- 31 -

ccc atc ccc atc cac tac tgc gcc ccc gcc ggc ttc gcc (SEQ ID NO:8).

oligo 3 reverse: gaa ctt ctt gtc ggc ggc gaa gcc ggc ggg (SEQ ID NO:9).

5 oligo 4 forward: gcg ccc ccg ccg gct tcg cca tcc tga agt gca acg aca aga agt tc (SEQ ID NO:10)

oligo 4: gcc gac aag aag ttc agc ggc aag ggc agc tgc aag aac gtg agc acc gtg cag tgc acc cac ggc atc cgg ccg gtg gtg agc acc cag ctc ctg ctg aac ggc agc ctg  
10 gcc gag gag gag gtg gtg atc cgc agc gag aac ttc acc gac aac gcc aag acc atc atc gtg cac ctg aat gag agc gtg cag atc (SEQ ID NO:11)

oligo 4 reverse (MluI): agt tgg gac gcg tgc agt tga tct gca cgc tct c (SEQ ID NO:12).

15 oligo 5 forward (MluI): gag agc gtg cag atc aac tgc acg cgt ccc (SEQ ID NO:13).

oligo 5: aac tgc acg cgt ccc aac tac aac aag cgc aag cgc atc cac atc ggc ccc ggg cgc gcc ttc tac acc acc aag aac atc atc ggc acc atc ctc cag gcc cac tgc aac atc  
20 tct aga (SEQ ID NO:14) .

oligo 5 reverse: gtc gtt cca ctt ggc tct aga gat gtt gca (SEQ ID NO:15).

oligo 6 forward: gca aca tct cta gag cca agt gga acg ac (SEQ ID NO:16).

25 oligo 6: gcc aag tgg aac gac acc ctg cgc cag atc gtg agc aag ctg aag gag cag ttc aag aac aag acc atc gtg ttc ac cag agc agc ggc ggc gac ccc gag atc gtg atg cac agc ttc aac tgc ggc ggc (SEQ ID NO:17).

oligo 6 reverse (EcoRI): gca gta gaa gaa ttc gcc  
30 gcc gca gtt ga (SEQ ID NO:18).

oligo 7 forward (EcoRI): tca act gcg gcg gcg aat tct tct act gc (SEQ ID NO:19).

oligo 7: ggc gaa ttc ttc tac tgc aac acc agc ccc ctg ttc aac agc acc tgg aac ggc aac aac acc tgg aac aac  
35 acc acc ggc agc aac aac aat att acc ctc cag tgc aag atc

- 32 -

aag cag atc atc aac atg tgg cag gag gtg ggc aag gcc atg  
tac gcc ccc ccc atc gag ggc cag atc cgg tgc agc agc (SEQ  
ID NO:20)

oligo 7 reverse: gca gac cgg tga tgt tgc tgc tgc  
5 acc gga tct ggc cct c (SEQ ID NO:21).

oligo 8 forward: cga ggg cca gat ccg gtg cag cag  
caa cat cac cgg tct g (SEQ ID NO:22).

oligo 8: aac atc acc ggt ctg ctg ctg acc cgc gac  
ggc ggc aag gac acc gac acc aac gac acc gaa atc ttc cgc  
10 ccc ggc ggc ggc gac atg cgc gac aac tgg aga tct gag ctg  
tac aag tac aag gtg gtg acg atc gag ccc ctg ggc gtg gcc  
ccc acc aag gcc aag cgc cgc gtg gtg cag cgc gag aag cgc  
(SEQ ID NO:23).

oligo 8 reverse (Not1): cgc ggg cgg ccg ctt tag  
15 cgc ttc tcg cgc tgc acc ac (SEQ ID NO:24).

The following oligonucleotides were used for the  
construction of the ratTHY-lenv gene.

oligo 1 forward (BamH1/Hind3): cgc ggg gga tcc  
aag ctt acc atg att cca gta ata agt (SEQ ID NO:25).

20 oligo 1: atg aat cca gta ata agt ata aca tta tta  
tta agt gta tta caa atg agt aga gga caa aga gta ata agt  
tta aca gca tct tta gta aat caa aat ttg aga tta gat tgt  
aga cat gaa aat aat aca aat ttg cca ata caa cat gaa ttt  
tca tta acg (SEQ ID NO:26).

25 oligo 1 reverse (EcoR1/Mlu1): cgc ggg gaa ttc acg  
cgt taa tga aaa ttc atg ttg (SEQ ID NO:27).

oligo 2 forward (BamH1/Mlu1): cgc gga tcc acg cgt  
gaa aaa aaa aaa cat (SEQ ID NO:28).

oligo 2: cgt gaa aaa aaa aaa cat gta tta agt gga  
30 aca tta gga gta cca gaa cat aca tat aga agt aga gta aat  
ttg ttt agt gat aga ttc ata aaa gta tta aca tta gca aat  
ttt aca aca aaa gat gaa gga gat tat atg tgt gag (SEQ ID  
NO:29).

oligo 2 reverse (EcoR1/Sac1): cgc gaa ttc gag ctc  
35 aca cat ata atc tcc (SEQ ID NO:30).

- 33 -

oligo 3 forward (BamH1/Sac1): cgc gga tcc gag ctc  
aga gta agt gga caa (SEQ ID NO:31).

oligo 3: ctc aga gta agt gga caa aat cca aca agt  
agt aat aaa aca ata aat gta ata aga gat aaa tta gta aaa  
5 tgt ga gga ata agt tta tta gta caa aat aca agt tgg tta  
tta tta tta tta tta agt tta agt ttt tta caa gca aca gat  
ttt ata agt tta tga (SEQ ID NO:32).

oligo 3 reverse (EcoR1/Not1): cgc gaa ttc gcg gcc  
gct tca taa act tat aaa atc (SEQ ID NO:33).

#### 10 Polymerase Chain Reaction

Short, overlapping 15 to 25 mer oligonucleotides  
annealing at both ends were used to amplify the long  
oligonuclotides by polymerase chain reaction (PCR).  
Typical PCR conditions were: 35 cycles, 55°C annealing  
15 temperature, 0.2 sec extension time. PCR products were  
gel purified, phenol extracted, and used in a subsequent  
PCR to generate longer fragments consisting of two  
adjacent small fragments. These longer fragments were  
cloned into a CDM7-derived plasmid containing a leader  
20 sequence of the CD5 surface molecule followed by a  
Nhe1/Pst1/Mlu1/EcoR1/BamH1 polylinker.

The following solutions were used in these  
reactions: 10x PCR buffer (500 mM KCl, 100 mM Tris HCl,  
pH 7.5, 8 mM MgCl<sub>2</sub>, 2 mM each dNTP). The final buffer  
25 was complemented with 10% DMSO to increase fidelity of  
the Taq polymerase.

#### Small scale DNA preparation

Transformed bacteria were grown in 3 ml LB  
cultures for more than 6 hours or overnight.  
30 Approximately 1.5 ml of each culture was poured into 1.5  
ml microfuge tubes, spun for 20 seconds to pellet cells  
and resuspended in 200 µl of solution I. Subsequently  
400 µl of solution II and 300 µl of solution III were  
added. The microfuge tubes were capped, mixed and spun  
35 for > 30 sec. Supernatants were transferred into fresh

- 34 -

tubes and phenol extracted once. DNA was precipitated by filling the tubes with isopropanol, mixing, and spinning in a microfuge for > 2 min. The pellets were rinsed in 70 % ethanol and resuspended in 50  $\mu$ l dH<sub>2</sub>O containing 10  $\mu$ l of RNase A. The following media and solutions were used in these procedures: LB medium (1.0 % NaCl, 0.5% yeast extract, 1.0% trypton); solution I (10 mM EDTA pH 8.0); solution II (0.2 M NaOH, 1.0% SDS); solution III (2.5 M KOAc, 2.5 M glacial acetic acid); phenol (pH adjusted to 6.0, overlaid with TE); TE (10 mM Tris HCl, pH 7.5, 1 mM EDTA pH 8.0).

#### Large scale DNA preparation

One liter cultures of transformed bacteria were grown 24 to 36 hours (MC1061p3 transformed with pCDM derivatives) or 12 to 16 hours (MC1061 transformed with pUC derivatives) at 37°C in either M9 bacterial medium (pCDM derivatives) or LB (pUC derivatives). Bacteria were spun down in 1 liter bottles using a Beckman J6 centrifuge at 4,200 rpm for 20 min. The pellet was resuspended in 40 ml of solution I. Subsequently, 80 ml of solution II and 40 ml of solution III were added and the bottles were shaken semivigorously until lumps of 2 to 3 mm size developed. The bottle was spun at 4,200 rpm for 5 min and the supernatant was poured through cheesecloth into a 250 ml bottle. Isopropanol was added to the top and the bottle was spun at 4,200 rpm for 10 min. The pellet was resuspended in 4.1 ml of solution I and added to 4.5 g of cesium chloride, 0.3 ml of 10 mg/ml ethidium bromide, and 0.1 ml of 1% Triton X100 solution. The tubes were spun in a Beckman J2 high speed centrifuge at 10,000 rpm for 5 min. The supernatant was transferred into Beckman Quick Seal ultracentrifuge tubes, which were then sealed and spun in a Beckman ultracentrifuge using a NVT90 fixed angle rotor at 80,000 rpm for > 2.5 hours. The band was extracted by visible light using a 1 ml

- 35 -

syringe and 20 gauge needle. An equal volume of  $\text{dH}_2\text{O}$  was added to the extracted material. DNA was extracted once with n-butanol saturated with 1 M sodium chloride, followed by addition of an equal volume of 10 M ammonium acetate/ 1 mM EDTA. The material was poured into a 13 ml snap tube which was then filled to the top with absolute ethanol, mixed, and spun in a Beckman J2 centrifuge at 10,000 rpm for 10 min. The pellet was rinsed with 70% ethanol and resuspended in 0.5 to 1 ml of  $\text{H}_2\text{O}$ . The DNA concentration was determined by measuring the optical density at 260 nm in a dilution of 1:200 ( $1 \text{ OD}_{260} = 50 \mu\text{g/ml}$ ).

The following media and buffers were used in these procedures: M9 bacterial medium (10 g M9 salts, 10 g casamino acids (hydrolyzed), 10 ml M9 additions, 7.5  $\mu\text{g/ml}$  tetracycline (500  $\mu\text{l}$  of a 15 mg/ml stock solution), 12.5  $\mu\text{g/ml}$  ampicillin (125  $\mu\text{l}$  of a 10 mg/ml stock solution); M9 additions (10 mM  $\text{CaCl}_2$ , 100 mM  $\text{MgSO}_4$ , 200  $\mu\text{g/ml}$  thiamine, 70% glycerol); LB medium (1.0 % NaCl, 0.5 % yeast extract, 1.0 % trypton); Solution I (10 mM EDTA pH 8.0); Solution II (0.2 M NaOH 1.0 % SDS); Solution III (2.5 M KOAc 2.5 M HOAc)

#### Sequencing

Synthetic genes were sequenced by the Sanger dideoxynucleotide method. In brief, 20 to 50  $\mu\text{g}$  double-stranded plasmid DNA were denatured in 0.5 M NaOH for 5 min. Subsequently the DNA was precipitated with 1/10 volume of sodium acetate (pH 5.2) and 2 volumes of ethanol and centrifuged for 5 min. The pellet was washed with 70% ethanol and resuspended at a concentration of 1  $\mu\text{g}/\mu\text{l}$ . The annealing reaction was carried out with 4  $\mu\text{g}$  of template DNA and 40 ng of primer in 1x annealing buffer in a final volume of 10  $\mu\text{l}$ . The reaction was heated to 65°C and slowly cooled to 37°C. In a separate tube 1  $\mu\text{l}$  of 0.1 M DTT, 2  $\mu\text{l}$  of labeling mix, 0.75  $\mu\text{l}$  of

- 36 -

dH<sub>2</sub>O, 1  $\mu$ l of [<sup>35</sup>S] dATP (10 uCi), and 0.25  $\mu$ l of Sequenase™ (12 U/ $\mu$ l) were added for each reaction. Five  $\mu$ l of this mix were added to each annealed primer-template tube and incubated for 5 min at room temperature. For each labeling reaction 2.5  $\mu$ l of each of the 4 termination mixes were added on a Terasaki plate and prewarmed at 37°C. At the end of the incubation period 3.5  $\mu$ l of labeling reaction were added to each of the 4 termination mixes. After 5 min, 4  $\mu$ l of stop solution were added to each reaction and the Terasaki plate was incubated at 80°C for 10 min in an oven. The sequencing reactions were run on 5% denaturing polyacrylamide gel. An acrylamide solution was prepared by adding 200 ml of 10x TBE buffer and 957 ml of dH<sub>2</sub>O to 100 g of acrylamide:bisacrylamide (29:1). 5% polyacrylamide 46% urea and 1x TBE gel was prepared by combining 38 ml of acrylamide solution and 28 g urea. Polymerization was initiated by the addition of 400  $\mu$ l of 10% ammonium peroxodisulfate and 60  $\mu$ l of TEMED. Gels were poured using silanized glass plates and sharktooth combs and run in 1x TBE buffer at 60 to 100 W for 2 to 4 hours (depending on the region to be read). Gels were transferred to Whatman blotting paper, dried at 80°C for about 1 hour, and exposed to x-ray film at room temperature. Typically exposure time was 12 hours. The following solutions were used in these procedures: 5x Annealing buffer (200 mM Tris HCl, pH 7.5, 100 mM MgCl<sub>2</sub>, 250 mM NaCl); Labelling Mix (7.5  $\mu$ M each dCTP, dGTP, and dTTP); Termination Mixes (80  $\mu$ M each dNTP, 50 mM NaCl, 8  $\mu$ M ddNTP (one each)); Stop solution (95% formamide, 20 mM EDTA, 0.05 % bromphenol blue, 0.05 % xylencyanol); 5x TBE (0.9 M Tris borate, 20 mM EDTA); Polyacrylamide solution (96.7 g polyacrylamide, 3.3 g bisacrylamide, 200 ml 1x TBE, 957 ml dH<sub>2</sub>O).

- 37 -

RNA isolation

Cytoplasmic RNA was isolated from calcium phosphate transfected 293T cells 36 hours post transfection and from vaccinia infected Hela cells 16 hours post infection essentially as described by Gilman. (Gilman Preparation of cytoplasmic RNA from tissue culture cells. In Current Protocols in Molecular Biology, Ausubel et al., eds., Wiley & Sons, New York, 1992). Briefly, cells were lysed in 400  $\mu$ l lysis buffer, nuclei were spun out, and SDS and proteinase K were added to 0.2% and 0.2 mg/ml respectively. The cytoplasmic extracts were incubated at 37°C for 20 min, phenol/chloroform extracted twice, and precipitated. The RNA was dissolved in 100  $\mu$ l buffer I and incubated at 37°C for 20 min. The reaction was stopped by adding 25  $\mu$ l stop buffer and precipitated again.

The following solutions were used in this procedure: Lysis Buffer (TRUSTEE containing with 50 mM Tris pH 8.0, 100 mM NaCl, 5 mM MgCl<sub>2</sub>, 0.5% NP40); Buffer I (TRUSTEE buffer with 10 mM MgCl<sub>2</sub>, 1 mM DTT, 0.5 U/ $\mu$ l placental RNase inhibitor, 0.1 U/ $\mu$ l RNase free DNase I); Stop buffer (50 mM EDTA 1.5 M NaOAc 1.0 % SDS).

Slot blot analysis

For slot blot analysis 10  $\mu$ g of cytoplasmic RNA was dissolved in 50  $\mu$ l dH<sub>2</sub>O to which 150  $\mu$ l of 10x SSC/18% formaldehyde were added. The solubilized RNA was then incubated at 65°C for 15 min and spotted onto with a slot blot apparatus. Radioactively labeled probes of 1.5 kb gp120IIIb and syngp120mn fragments were used for hybridization. Each of the two fragments was random labeled in a 50  $\mu$ l reaction with 10  $\mu$ l of 5x oligo-labeling buffer, 8  $\mu$ l of 2.5 mg/ml BSA, 4  $\mu$ l of  $\alpha$ [<sup>32</sup>P]-dCTP (20 uCi/ $\mu$ l; 6000 Ci/mmol), and 5 U of Klenow fragment. After 1 to 3 hours incubation at 37°C 100  $\mu$ l of TRUSTEE were added and unincorporated  $\alpha$ [<sup>32</sup>P]-dCTP was

- 38 -

eliminated using G50 spin column. Activity was measured in a Beckman beta-counter, and equal specific activities were used for hybridization. Membranes were pre-hybridized for 2 hours and hybridized for 12 to 24 hours at 42°C with  $0.5 \times 10^6$  cpm probe per ml hybridization fluid. The membrane was washed twice (5 min) with washing buffer I at room temperature, for one hour in washing buffer II at 65°C, and then exposed to x-ray film. Similar results were obtained using a 1.1 kb NotI/SfiI fragment of pCDM7 containing the 3 untranslated region. Control hybridizations were done in parallel with a random-labeled human beta-actin probe. RNA expression was quantitated by scanning the hybridized nitrocellulose membranes with a Magnetic Dynamics phosphorimager.

The following solutions were used in this procedure:

5x Oligo-labeling buffer (250 mM Tris HCl, pH 8.0, 25 mM  $MgCl_2$ , 5 mM  $\beta$ -mercaptoethanol, 2 mM dATP, 2 mM dGTP, mM dTTP, 1 M Hepes pH 6.6, 1 mg/ml hexanucleotides [dNTP]6); Hybridization Solution (\_\_\_ M sodium phosphate, 250 mM NaCl, 7% SDS, 1 mM EDTA, 5% dextrane sulfate, 50% formamide, 100  $\mu$ g/ml denatured salmon sperm DNA); Washing buffer I (2x SSC, 0.1% SDS); Washing buffer II (0.5x SSC, 0.1 % SDS); 20x SSC (3 M NaCl, 0.3 M  $Na_3$ citrate, pH adjusted to 7.0).

#### Vaccinia recombination

Vaccinia recombination used a modification of the of the method described by Romeo and Seed (Romeo and Seed, Cell, 64: 1037, 1991). Briefly, CV1 cells at 70 to 90% confluency were infected with 1 to 3  $\mu$ l of a wild-type vaccinia stock WR ( $2 \times 10^8$  pfu/ml) for 1 hour in culture medium without calf serum. After 24 hours, the cells were transfected by calcium phosphate with 25  $\mu$ g TKG plasmid DNA per dish. After an additional 24 to 48



- 39 -

hours the cells were scraped off the plate, spun down, and resuspended in a volume of 1 ml. After 3 freeze/thaw cycles trypsin was added to 0.05 mg/ml and lysates were incubated for 20 min. A dilution series of 10, 1 and 0.1  $\mu$ l of this lysate was used to infect small dishes (6 cm) of CV1 cells, that had been pretreated with 12.5  $\mu$ g/ml mycophenolic acid, 0.25 mg/ml xanthin and 1.36 mg/ml hypoxanthine for 6 hours. Infected cells were cultured for 2 to 3 days, and subsequently stained with the monoclonal antibody NEA9301 against gp120 and an alkaline phosphatase conjugated secondary antibody. Cells were incubated with 0.33 mg/ml NBT and 0.16 mg/ml BCIP in AP-buffer and finally overlaid with 1% agarose in PBS. Positive plaques were picked and resuspended in 100  $\mu$ l Tris pH 9.0. The plaque purification was repeated once. To produce high titer stocks the infection was slowly scaled up. Finally, one large plate of Hela cells was infected with half of the virus of the previous round. Infected cells were detached in 3 ml of PBS, lysed with a Dounce homogenizer and cleared from larger debris by centrifugation. VPE-8 recombinant vaccinia stocks were kindly provided by the AIDS repository, Rockville, MD, and express HIV-1 IIIB gp120 under the 7.5 mixed early/late promoter (Earl et al., *J. Virol.*, 65:31, 1991). In all experiments with recombinant vaccinia cells were infected at a multiplicity of infection of at least 10.

The following solution was used in this procedure:  
AP buffer (100 mM Tris HCl, pH 9.5, 100 mM NaCl, 5 mM  
MgCl<sub>2</sub>)

#### Cell culture

The monkey kidney carcinoma cell lines CV1 and Cos7, the human kidney carcinoma cell line 293T, and the human cervix carcinoma cell line Hela were obtained from the American Tissue Typing Collection and were maintained

- 40 -

in supplemented IMDM. They were kept on 10 cm tissue culture plates and typically split 1:5 to 1:20 every 3 to 4 days. The following medium was used in this procedure:

- 5 Supplemented IMDM (90% Iscove's modified Dulbecco Medium, 10% calf serum, iron-complemented, heat inactivated 30 min 56°C, 0.3 mg/ml L-glutamine, 25 µg/ml gentamycin 0.5 mM β-mercaptoethanol (pH adjusted with 5 M NaOH, 0.5 ml)).

10 Transfection

Calcium phosphate transfection of 293T cells was performed by slowly adding and under vortexing 10 µg plasmid DNA in 250 µl 0.25 M CaCl<sub>2</sub> to the same volume of 2x HEBS buffer while vortexing. After incubation for 10  
15 to 30 min at room temperature the DNA precipitate was added to a small dish of 50 to 70% confluent cells. In cotransfection experiments with rev, cells were transfected with 10 µg gp120IIIb, gp120IIIbrre, syngp120mrre or rTHY-lenveglrre and 10 µg of pCMVrev or  
20 CDM7 plasmid DNA.

The following solutions were used in this procedure: 2x HEBS buffer (280 mM NaCl, 10 mM KCl, 1.5 mM sterile filtered); 0.25 mM CaCl<sub>2</sub> (autoclaved).

Immunoprecipitation

- 25 After 48 to 60 hours medium was exchanged and cells were incubated for additional 12 hours in Cys/Met-free medium containing 200 µCi of <sup>35</sup>S-translabel. Supernatants were harvested and spun for 15 min at 3000 rpm to remove debris. After addition of protease  
30 inhibitors leupeptin, aprotinin and PMSF to 2.5 µg/ml, 50 µg/ml, 100 µg/ml respectively, 1 ml of supernatant was incubated with either 10 µl of packed protein A sepharose alone (rTHY-lenveglrre) or with protein A sepharose and 3 µg of a purified CD4/immunoglobulin fusion protein  
35 (kindly provided by Behring) (all gp120 constructs) at

- 41 -

4°C for 12 hours on a rotator. Subsequently the protein A beads were washed 5 times for 5 to 15 min each time. After the final wash 10 µl of loading buffer containing was added, samples were boiled for 3 min and applied on 5 7% (all gp120 constructs) or 10% (rTHY-1enveglirre) SDS polyacrylamide gels (TRIS pH 8.8 buffer in the resolving, TRIS pH 6.8 buffer in the stacking gel, TRIS-glycin running buffer, Maniatis et al. 1989). Gels were fixed in 10% acetic acid and 10 % methanol, incubated with 10 Amplify for 20 min, dried and exposed for 12 hours.

The following buffers and solutions were used in this procedure: Wash buffer (100 mM Tris, pH 7.5, 150 mM NaCl, 5 mM CaCl<sub>2</sub>, 1% NP-40); 5x Running Buffer (125 mM Tris, 1.25 M Glycin, 0.5% SDS); Loading buffer (10 % 15 glycerol, 4% SDS, 4% β-mercaptoethanol, 0.02 % bromphenol blue).

#### Immunofluorescence

293T cells were transfected by calcium phosphate coprecipitation and analyzed for surface THY-1 expression 20 after 3 days. After detachment with 1 mM EDTA/PBS, cells were stained with the monoclonal antibody OX-7 in a dilution of 1:250 at 4°C for 20 min, washed with PBS and subsequently incubated with a 1:500 dilution of a FITC-conjugated goat anti-mouse immunoglobulin antiserum. 25 Cells were washed again, resuspended in 0.5 ml of a fixing solution, and analyzed on a EPICS XL cytofluorometer (Coulter).

The following solutions were used in this procedure:

30 PBS (137 mM NaCl, 2.7 mM KCl, 4.3 mM Na<sub>2</sub>HPO<sub>4</sub>, 1.4 mM KH<sub>2</sub>PO<sub>4</sub>, pH adjusted to 7.4); Fixing solution (2% formaldehyde in PBS).

#### ELISA

The concentration of gp120 in culture supernatants 35 was determined using CD4-coated ELISA plates and goat

- 42 -

anti-gp120 antisera in the soluble phase. Supernatants of 293T cells transfected by calcium phosphate were harvested after 4 days, spun at 3000 rpm for 10 min to remove debris and incubated for 12 hours at 4°C on the plates. After 6 washes with PBS 100 µl of goat anti-gp120 antisera diluted 1:200 were added for 2 hours. The plates were washed again and incubated for 2 hours with a peroxidase-conjugated rabbit anti-goat IgG antiserum 1:1000. Subsequently the plates were washed and incubated for 30 min with 100 µl of substrate solution containing 2 mg/ml o-phenylenediamine in sodium citrate buffer. The reaction was finally stopped with 100 µl of 4 M sulfuric acid. Plates were read at 490 nm with a Coulter microplate reader. Purified recombinant gp120IIIb was used as a control. The following buffers and solutions were used in this procedure: Wash buffer (0.1% NP40 in PBS); Substrate solution (2 mg/ml o-phenylenediamine in sodium citrate buffer).

#### Green Fluorescent Protein

The efficacy of codon replacement for gp120 suggests that replacing non-preferred codons with less preferred codons or preferred codons (and replacing less preferred codons with preferred codons) will increase expression in mammalian cells of other proteins, e.g., other eukaryotic proteins.

The green fluorescent protein (GFP) of the jellyfish *Aequorea victoria* (Ward, Photochem. Photobiol. 4:1, 1979; Prasher et al., Gene 111:229, 1992; Cody et al., Biochem. 32:1212, 1993) has attracted attention recently for its possible utility as a marker or reporter for transfection and lineage studies (Chalfie et al., Science 263:802, 1994).

Examination of a codon usage table constructed from the native coding sequence of GFP showed that the GFP codons favored either A or U in the third position.

- 43 -

The bias in this case favors A less than does the bias of gp120, but is substantial. A synthetic gene was created in which the natural GFP sequence was re-engineered in much the same manner as for gp120. In addition, the translation initiation sequence of GFP was replaced with sequences corresponding to the translational initiation consensus. The expression of the resulting protein was contrasted with that of the wild type sequence, similarly engineered to bear an optimized translational initiation consensus (FIG. 10, panel B and FIG. 10, panel C). In addition, the effect of inclusion of the mutation Ser 65→Thr, reported to improve excitation efficiency of GFP at 490 nm and hence preferred for fluorescence microscopy (Heim et al., Nature 373:663,1995), was examined (FIG. 10, panel D). Codon engineering conferred a significant increase in expression efficiency (an concomitant percentage of cells apparently positive for transfection), and the combination of the Ser 65→Thr mutation and codon optimization resulted in a DNA segment encoding a highly visible mammalian marker protein (FIG. 10, panel D).

The above-described synthetic green fluorescent protein coding sequence was assembled in a similar manner as for gp120 from six fragments of approximately 120 bp each, using a strategy for assembly that relied on the ability of the restriction enzymes BsaI and BbsI to cleave outside of their recognition sequence. Long oligonucleotides were synthesized which contained portions of the coding sequence for GFP embedded in flanking sequences encoding EcoRI and BsaI at one end, and BamHI and BbsI at the other end. Thus, each oligonucleotide has the configuration EcoRI/BsaI/GFP fragment/BbsI/BamHI. The restriction site ends generated by the BsaI and BbsI sites were designed to yield compatible ends that could be used to join adjacent GFP

- 44 -

fragments. Each of the compatible ends were designed to be unique and non-selfcomplementary. The crude synthetic DNA segments were amplified by PCR, inserted between EcoRI and BamHI in pUC9, and sequenced. Subsequently the intact coding sequence was assembled in a six fragment ligation, using insert fragments prepared with BsaI and BbsI. Two of six plasmids resulting from the ligation bore an insert of correct size, and one contained the desired full length sequence. Mutation of Ser65 to Thr was accomplished by standard PCR based mutagenesis, using a primer that overlapped a unique BssSI site in the synthetic GFP.

Codon optimization as a strategy for improved expression in mammalian cells

The data presented here suggest that coding sequence re-engineering may have general utility for the improvement of expression of mammalian and non-mammalian eukaryotic genes in mammalian cells. The results obtained here with three unrelated proteins: HIV gp120, the rat cell surface antigen Thy-1 and green fluorescent protein from *Aequorea victoria*, suggest that codon optimization may prove to be a fruitful strategy for improving the expression in mammalian cells of a wide variety of eukaryotic genes.

Use

The synthetic genes of the invention are useful for expressing the a protein normally expressed in mammalian cells in cell culture (e.g. for commercial production of human proteins such as hGH, TPA, Factor VII, and Factor IX). The synthetic genes of the invention are also useful for gene therapy.

Synthetic GFP genes can be used in any application in which a native GFP gene or other reporter gene can be used. A synthetic GFP gene which employs more preferred

- 45 -

codons than the native GFP gene can be the basis of a highly sensitive reporter system. Such a system can be used, e.g., to analyze the influence of particular promoter elements or trans-acting factors on gene  
5 expression. Thus, the synthetic GFP gene can be used in much the same fashion as other reporters, e.g.,  $\beta$ -galactosidase, has been used.

- 46 -

## SEQUENCE LISTING

## (1) GENERAL INFORMATION:

- (i) APPLICANT: THE GENERAL HOSPITAL CORPORATION
- (ii) TITLE OF INVENTION: HIGH LEVEL EXPRESSION OF PROTEINS
- (iii) NUMBER OF SEQUENCES: 40
- (iv) CORRESPONDENCE ADDRESS:
  - (A) ADDRESSEE: Fish & Richardson P.C.
  - (B) STREET: 225 Franklin Street
  - (C) CITY: Boston
  - (D) STATE: Massachusetts
  - (E) COUNTRY: U.S.A.
  - (F) ZIP: 02110-2804
- (v) COMPUTER READABLE FORM:
  - (A) MEDIUM TYPE: Floppy disk
  - (B) COMPUTER: IBM PC compatible
  - (C) OPERATING SYSTEM: PC-DOS/MS-DOS
  - (D) SOFTWARE: PatentIn Release #1.0, Version #1.30B
- (vi) CURRENT APPLICATION DATA:
  - (A) APPLICATION NUMBER: PCT/US96/-----
  - (B) FILING DATE: -SEP-1996
- (vii) PRIOR APPLICATION DATA:
  - (A) APPLICATION NUMBER: 08/532,390
  - (B) FILING DATE: 22-SEP-1995
- (vii) PRIOR APPLICATION DATA:
  - (A) APPLICATION NUMBER: 08/324,243
  - (B) FILING DATE: 19-SEP-1994
- (viii) ATTORNEY/AGENT INFORMATION:
  - (A) NAME: LECH, KAREN F.
  - (B) REGISTRATION NUMBER: 35,238
  - (C) REFERENCE/DOCKET NUMBER: 00786/294001
- (ix) TELECOMMUNICATION INFORMATION:
  - (A) TELEPHONE: (617) 542-5070
  - (B) TELEFAX: (617) 542-8906
  - (C) TELEX: 200154

## (2) INFORMATION FOR SEQ ID NO:1:

- (i) SEQUENCE CHARACTERISTICS:
  - (A) LENGTH: 24 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

CGCGGGCTAG CCACCGAGAA GCTG



- 47 -

## (2) INFORMATION FOR SEQ ID NO:2:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 196 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:

ACCGAGAAGC TGTGGGTGAC CGTGTACTAC GCGTGCCCG TGTGGAAGAG AGGCCACCAC	60
CACCCTGTTC TCGCCAGCG ACGCCAAGGC GTACGACACC GAGGTGCACA ACGTGTGGGC	120
CACCCAGGCG TCGTGCCCA CCGACCCAA CCCCCAGGAG GTGGAGCTCG TGAACGTGAC	180
CGAGAACTTC AACATG	196

## (2) INFORMATION FOR SEQ ID NO:3:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 34 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

CCACCATGTT GTTCTTCCAC ATGTTGAAGT TCTC	34
---------------------------------------	----

## (2) INFORMATION FOR SEQ ID NO:4:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 33 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:4:

GACCGAGAAC TTCAACATGT GGAAGAACAA CAT	33
--------------------------------------	----

## (2) INFORMATION FOR SEQ ID NO:5:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 192 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

- 48 -

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:5:

TGGAAGAACA ACATGGTGGA GCAGATGCAT GAGGACATCA TCAGCCTGTG GGACCAGAGC	60
CTGAAGCCCT GCGTGAAGCT GACCCCTGT GCGTGACCTG AACTGCACCG ACCTGAGGAA	120
CACCACCAAC ACCAACACAG CACCGCCAAC AACAAACAGCA ACAGCGAGGG CACCATCAAG	180
GGCGGCGAGA TG	192

## (2) INFORMATION FOR SEQ ID NO:6:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 33 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:6:

GTGAAGCTG CAGTTCTTCA TCTCGCCGCC CTT	33
-------------------------------------	----

## (2) INFORMATION FOR SEQ ID NO:7:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 31 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7:

GAAGAACTGC AGCTTCAACA TCACCACCAG C	31
------------------------------------	----

## (2) INFORMATION FOR SEQ ID NO:8:

- (i) SEQUENCE CHARACTERISTICS:
- (A) LENGTH: 195 base pairs
  - (B) TYPE: nucleic acid
  - (C) STRANDEDNESS: single
  - (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8:

AACATCACCA CCAGCATCCG CGACAAGATG CAGAAGGAGT ACGCCCTGCT GTACAAGCTG	60
GATATCGTGA GCATCGACAA CGACAGCACC AGCTACCGCC TGATCTCCTG CAACACCAGC	120
GTGATCACCC AGGCCTGCCC CAAGATCAGC TTCGAGCCCA TCCCCATCCA CTACTGCGCC	180
CCCGCCGGCT TCGCC	195

- 49 -

## (2) INFORMATION FOR SEQ ID NO:9:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 30 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9:

GAACTTCTTG TCGGCGGCGA AGCCGGCGGG

30

## (2) INFORMATION FOR SEQ ID NO:10:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 47 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10:

GCGCCCCCGC CGGCTTCGCC ATCCTGAAGT GCAACGACAA GAAGTTC

47

## (2) INFORMATION FOR SEQ ID NO:11:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 198 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11:

GCCGACAAGA AGTTCAGCGG CAAGGGCAGC TGCAAGAACG TGAGCACCGT GCAGTGCACC 60  
CACGGCATCC GGCCGGTGGT GAGCACCCAG CTCCTGCTGA ACGGCAGCCT GGCCGAGGAG 120  
GAGGTGGTGA TCCGCAGCGA GAACTTCACC GACAACGCCA AGACCATCAT CGTGCACCTG 180  
AATGAGAGCG TGCAGATC 198

## (2) INFORMATION FOR SEQ ID NO:12:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 34 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

- 50 -

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:12:

AGTTGGGACG CGTGCAGTTG ATCTGCACGC TCTC

34

(2) INFORMATION FOR SEQ ID NO:13:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:13:

GAGAGCGTGC AGATCAACTG CACGCGTCCC

30

(2) INFORMATION FOR SEQ ID NO:14:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 120 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:14:

AACTGCACGC GTCCCAACTA CAACAAGCGC AAGCGCATCC ACATCGGCCC CGGGCGCGCC  
TTCTACACCA CCAAGAACAT CATCGGCACC ATCCTCCAGG CCCACTGCAA CATCTCTAGA

60

120

(2) INFORMATION FOR SEQ ID NO:15:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:15:

GTCGTTCCAC TTGGCTCTAG AGATGTTGCA

30

(2) INFORMATION FOR SEQ ID NO:16:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 29 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

- 51 -

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16:

GCAACATCTC TAGAGCCAAG TGGAACGAC

29

## (2) INFORMATION FOR SEQ ID NO:17:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 131 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:17:

GCCAAGTGGA ACGACACCCT GCGCCAGATC GTGAGCAAGC TGAAGGAGCA GTTCAAGAAC 60

AAGACCATCG TGTTCAACCAG AGCAGCGGCG GCGACCCCGA GATCGTGATG CACAGCTTCA 120

ACTGCGGCGG C 131

## (2) INFORMATION FOR SEQ ID NO:18:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 29 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18:

GCAGTAGAAG AATTCGCCGC CGCAGTTGA

29

## (2) INFORMATION FOR SEQ ID NO:19:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 29 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19:

TCAACTGCGG CGGCGAATTC TTCTACTGC

29

## (2) INFORMATION FOR SEQ ID NO:20:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 195 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

- 52 -

(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20:

GGCGAATTCT TCTACTGCAA CACCAGCCCC CTGTTCAACA GCACCTGGAA CGGCAACAAC 60  
ACCTGGAACA ACACCACCGG CAGCAACAAC AATATTACCC TCCAGTGCAA GATCAAGCAG 120  
ATCATCAACA TGTGGCAGGA GGTGGGCAAG GCCATGTACG CCCCCCCCAT CGAGGGCCAG 180  
ATCCGGTGCA GCAGC 195

(2) INFORMATION FOR SEQ ID NO:21:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 40 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:21:

GCAGACCGGT GATGTTGCTG CTGCACCGGA TCTGGCCCTC 40

(2) INFORMATION FOR SEQ ID NO:22:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 40 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:22:

CGAGGGCCAG ATCCGGTGCA GCAGCAACAT CACCGGTCTG 40

(2) INFORMATION FOR SEQ ID NO:23:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 242 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

- 53 -

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:23:

AACATCACCG GTCTGCTGCT GCTGCTGACC CGGACGGCGG CAAGGACACC GACACCAACG 60  
ACACCGAAAT CTTCCGCGAC GCGGCAAGG ACACCAACGA CACCGAAATC TTCCGCCCCG 120  
GCGGCGGCGA CATGCGCGAC AACTGGAGAT CTGAGCTGTA CAAGTACAAG GTGGTGACGA 180  
TCGAGCCCCT GGGCGTGGCC CCCACCAAGG CCAAGCGCGC GGTGGTGACG CGCGAGAAGC 240  
GC 242

## (2) INFORMATION FOR SEQ ID NO:24:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 38 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:24:

CGCGGGCGGC CGCTTTAGCG CTTCTCGCGC TGCACCAC 38

## (2) INFORMATION FOR SEQ ID NO:25:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 39 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

CGCGGGGGAT CCAAGCTTAC CATGATTCCA GTAATAAGT 39

## (2) INFORMATION FOR SEQ ID NO:26:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 165 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:26:

ATGAATCCAG TAATAAGTAT AACATTATTA TTAAGTGTAT TACAAATGAG TAGAGGACAA 60  
AGAGTAATAA GTTAAACAGC ATCTTTAGTA AATCAAATT TGAGATTAGA TTGTAGACAT 120  
GAAATAATA CAAATTGACC AATACAACAT GAATTTTCAT TAACG 165

- 54 -

## (2) INFORMATION FOR SEQ ID NO:27:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 36 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:27:

CGCGGGGAAT TCACGCGTTA ATGAAAATTC ATGTTG

36

## (2) INFORMATION FOR SEQ ID NO:28:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 30 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:28:

CGCGGATCCA CGCGTGAAAA AAAAAACAT

30

## (2) INFORMATION FOR SEQ ID NO:29:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 149 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29:

CGTGAAAAAA AAAACATGT ATTAAGTGGA ACATTAGGAG TACCAGAACA TACATATAGA

60

AGTAGAGTAA TTTGTTTAGT GATAGATTCA TAAAAGTATT AACATTAGCA AATTTTACAA

120

CAAAAGATGA AGGAGATTAT ATGTGTGAG

149

## (2) INFORMATION FOR SEQ ID NO:30:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 30 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear



- 55 -

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:30:

CGCGAATTCG AGCTCACACA TATAATCTCC

30

## (2) INFORMATION FOR SEQ ID NO:31:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 30 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:31:

CGCGGATCCG AGCTCAGAGT AAGTGGACAA

30

## (2) INFORMATION FOR SEQ ID NO:32:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 170 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:32:

CTCAGAGTAA GTGGACAAA TCCAACAAGT AGTAATAAAA CAATAAATGT AATAAGAGAT 60

AAATTAGTAA AATGTGAGGA ATAAGTTTAT TAGTACAAAA TACAAGTTGG TTATTATTAT 120

TATTATTAAG TTTAAGTTTT TTACAAGCAA CAGATTTTAT AAGTTTATGA 170

## (2) INFORMATION FOR SEQ ID NO:33:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 36 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:33:

CGCGAATTCG CGGCCGCTTC ATAAACTTAT AAAATC

36

## (2) INFORMATION FOR SEQ ID NO:34:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 1632 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single

- 56 -

(D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:34:

CTCGAGATCC ATTGTGCTCT AAAGGAGATA CCCGGCCAGA CACCCTCACC TGCGGTGCCC	60
AGCTGCCCAG GCTGAGGCAA GAGAAGGCCA GAAACCATGC CCATGGGGTC TCTGCAACCG	120
CTGGCCACCT TGTACCTGCT GGGGATGCTG GTCGCTTCCG TGCTAGCCAC CGAGAAGCTG	180
TGGGTGACCG TGTACTACGG CGTGCCCGTG TGGAAGGAGG CCACCACCAC CCTGTTCTGC	240
GCCAGCGACG CCAAGGCGTA CGACACCGAG GTGCACAACG TGTGGGCCAC CCAGGCGTGC	300
GTGCCCACCG ACCCCAACCC CCAGGAGGTG GAGCTCGTGA ACGTGACCGA GAACTTCAAC	360
ATGTGGAAGA ACAACATGGT GGAGCAGATG CATGAGGACA TCATCAGCCT GTGGGACCAG	420
AGCCTGAAGC CCTGCGTGAA GCTGACCCCC CTGTGCGTGA CCCTGAACTG CACCGACCTG	480
AGGAACACCA CCAACACCAA CAACAGCACC GCCAACAACA ACAGCAACAG CGAGGGCACC	540
ATCAAGGGCG GCGAGATGAA CAACTGCAGC TTCAACATCA CCACCAGCAT CCGCGACAAG	600
ATGCAGAAGG AGTACGCCCT GCTGTACAAG CTGGATATCG TGAGCATCGA CAACGACAGC	660
ACCAGCTACC GCCTGATCTC CTGCAACACC AGCGTGATCA CCCAGGCCTG GCCCAAGATC	720
AGCTTCGAGC CCATCCCCAT CCACTACTGC GCCCCCGCCG GCTTCGCCAT CCTGAAGTGC	780
AACGACAAGA AGTTCAGCGG CAAGGGCAGC TGCAAGAACG TGAGCACCGT GCAGTGCACC	840
CACGGCATCC GGCCGGTGGT GAGCACCCAG CTCCTGCTGA ACGGCAGCCT GGCCGAGGAG	900
GAGGTGGTGA TCCGCAGCGA GAACTTCACC GACAACGCCA AGACCATCAT CGTGCACCTG	960
AATGAGAGCG TGCAGATCAA CTGCACGCGT CCCAACTACA ACAAGCGCAA GCGCATCCAC	1020
ATCGGCCCCG GCGCGCCTT CTACACCACC AAGAACATCA TCGGCACCAT CCGCCAGGCC	1080
CACTGCAACA TCTCTAGAGC CAAGTGGAAC GACACCCTGC GCCAGATCGT GAGCAAGCTG	1140
AAGGAGCAGT TCAAGAACAA GACCATCGTG TTCAACCAGA GCAGCGGCGG CGACCCCGAG	1200
ATCGTGATGC ACAGCTTCAA CTGCGGCGGC GAATTCTTCT ACTGCAACAC CAGCCCCCTG	1260
TTCAACAGCA CCTGGAACGG CAACAACACC TGAACAACA CCACCGGCAG CAACAACAAT	1320
ATTACCCTCC AGTGCAAGAT CAAGCAGATC ATCAACATGT GGCAGGAGGT GGGCAAGGCC	1380
ATGTACGCC CCCCATCGA GGGCCAGATC CGGTGCAGCA GCAACATCAC CGGTCTGCTG	1440
CTGACCCGCG ACGGCGGCAA GGACACCGAC ACCAACGACA CCGAAATCTT CCGCCCCGGC	1500
GGCGGCGACA TGCGCGACAA CTGGAGATCT GAGCTGTACA AGTACAAGGT GGTGACGATC	1560
GAGCCCCTGG GCGTGGCCCC CACCAAGGCC AAGCGCCGCG TGGTGCAGCG CGAGAAGCGC	1620
TAAAGCGGCC GC	1632

- 57 -

## (2) INFORMATION FOR SEQ ID NO:35:

## (i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 2481 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:35:

ACCGAGAAGC TGTGGGTGAC CGTGTACTAC GCGGTGCCCCG TGTGGAAGGA GGCCACCACC	60
ACCCTGTTCT GCGCCAGCGA CGCCAAGGCG TACGACACCG AGGTGCACAA CGTGTGGGCC	120
ACCCAGGCGT GCGTGCCAC CGACCCCAAC CCCAGGAGG TGGAGCTCGT GAACGTGACC	180
GAGAACTTCA ACATGTGGAA GAACAACATG CTGGAGCAGA TGCATGAGGA CATCATCAGC	240
CTGTGGGACC AGAGCCTGAA GCCCTGCGTG AAGCTGACCC CCCTGTGCGT GACCCTGAAC	300
TGCACCGACC TGAGGAACAC CACCAACACC AACAACAGCA CCGCCAACAA CAACAGCAAC	360
AGCGAGGGCA CCATCAAGGG CGGCGAGATG AAGAACTGCA GCTTCAACAT CACCACCAGC	420
ATCCGCGACA AGATGCAGAA GGAGTACGCC CTGCTGTACA AGCTGGATAT CGTGAGCATC	480
CACAACGACA GCACCAGCTA CCGCCTGATC TCCTGCAACA CCAGCGTGAT CACCCAGGCC	540
TGCCCCAAGA TCAGCTTCGA GCCCATCCCC ATCCACTACT GCGCCCCCGC CGGCTTCGCC	600
ATCCTGAAGT GCAACGACAA GAAGTTCAGC GGCAAGGGCA GCTGCAAGAA CGTGACCACC	660
GTGCAGTGCA CCCACGGCAT CCGGCCGGTG GTGAGCACCC AGCTCCTGCT GAACGGCAGC	720
CTGGCCGAGG AGGAGGTGGT GATCCGCAGC GAGAACTTCA CCGACAACGC CAAGACCATC	780
ATCGTGCACC TGAATGAGAG CGTGCAGATC AACTGCACGC GTCCCAACTA CAACAAGCGC	840
AAGCGCATCC ACATCGGCC CGGGCGCGCC TTCTACACCA CCAAGAACAT CATCGGCACC	900
ATCCGCCAGG CCCACTGCAA CATCTCTAGA GCCAAGTGGA ACGACACCCT GCGCCAGATC	960
GTGAGCAAGC TGAAGGAGCA GTTCAAGAAC AAGACCATCG TGTTCACCA GAGCAGCGGC	1020
GGCGACCCCG AGATCGTGAT GCACAGCTT CACTGCGGCG GCGAATTCTT CTACTGCAAC	1080
ACCAGCCCCC TGTTC AACAG CACCTGGAAC GGCAACAACA CCTGGAACAA CACCACCGGC	1140
AGCAACAACA ATATTACCCT CCAAGTGCAAG ATCAAGCAGA TCATCAACAT GTGGCAGGAG	1200
GTGGGCAAGG CCATGTACGC CCCCCCATC GAGGGCCAGA TCCGGTGCAG CAGCAACATC	1260
ACCGGTCTGC TGCTGACCCG CGACGGCGGC AAGGACACCG ACACCAACGA CACCGAAATC	1320
TTCCGCCCCG GCGGCGGCGA CATGCGCGAC AACTGGAGAT CTGAGCTGTA CAAGTACAAG	1380
GTGGTGACGA TCGAGCCCCT GGGCGTGGCC CCCACCAAGG CCAAGCGCCG CGTGGTGCAG	1440
CGCGAGAAGC GGGCCGCCAT CGGCGCCCTG TTCCTGGGCT TCCTGGGGGC GGCGGGCAGC	1500

- 58 -

ACCATGGGGG	CCGCCAGCGT	GACCCTGACC	GTGCAGGCCC	GCCTGCTCCT	GAGCGGCATC	1560
GTGCAGCAGC	AGAACAACCT	CCTCCGCGCC	ATCGAGGCCC	AGCAGCATAT	GCTCCAGCTC	1620
ACCGTGTGGG	GCATCAAGCA	GCTCCAGGCC	CGCGTGCTGG	CCGTGGAGCG	CTACCTGAAG	1680
GACCAGCAGC	TCCTGGGCTT	CTGGGGCTGC	TCCGGCAAGC	TGATCTGCAC	CACCACGGTA	1740
CCCTGGAACG	CCTCCTGGAG	CAACAAGAGC	CTGGACGACA	TCTGGAACAA	CATGACCTGG	1800
ATGCAGTGGG	AGCCGCGAGAT	CGATAACTAC	ACCAGCCTGA	TCTACAGCCT	GCTGGAGAAG	1860
AGCCAGACCC	AGCAGGAGAA	GAACGAGCAG	GAGCTGCTGG	AGCTGGACAA	CTGGGCGAGC	1920
CTGTGGAACT	GGTTCGACAT	CACCAACTGG	CTGTGGTACA	TCAAAATCTT	CATCATGATT	1980
GTGGGCGGCC	TGGTGGGCCT	CCGCATCGTG	TTCGCCGTGC	TGAGCATCGT	GAACCGCGTG	2040
CGCCAGGGCT	ACAGCCCCCT	GAGCCTCCAG	ACCCGGCCCC	CCGTGCCGCG	CGGGCCCCGAC	2100
CGCCCCGAGG	GCATCGAGGA	GGAGGGCGGC	GAGCGCGACC	GCGACACCAG	CGGCAGGCTC	2160
GTGCACGGCT	TCCTGGCGAT	CATCTGGGTC	GACCTCCGCA	GCCTGTTCTT	GTTCAGCTAC	2220
CACCACCGCG	ACCTGCTGCT	GATCGCCGCC	CGCATCGTGG	AACTCCTAGG	CCGCCGCGGC	2280
TGGGAGGTGC	TGAAGTACTG	GTGGAACCTC	CTCCAGTATT	GGAGCCAGGA	GCTGAAGTCC	2340
AGCGCCGTGA	GCCTGCTGAA	CGCCACCGCC	ATCGCCGTGG	CCGAGGGCAC	CGACCGCGTG	2400
ATCGAGGTGC	TCCAGAGGGC	CGGGAGGGCG	ATCCTGCACA	TCCCCACCCG	CATCCGCCAG	2460
GGGCTCGAGA	GGGCGCTGCT	G				2481

## (2) INFORMATION FOR SEQ ID NO:36:

- (i) SEQUENCE CHARACTERISTICS:  
 (A) LENGTH: 486 base pairs  
 (B) TYPE: nucleic acid  
 (C) STRANDEDNESS: single  
 (D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:36:

ATGAATCCAG	TAATAAGTAT	AACATTATTA	TTAAGTGTAT	TACAAATGAG	TAGAGGACAA	60
AGAGTAATAA	GTTTAACAGC	ATGTTTAGTA	AATCAAAATT	TGAGATTAGA	TTGTAGACAT	120
GAAAATAATA	CACCTTTGCC	AATACAACAT	GAATTTTCAT	TAACGCGTGA	AAAAAAAAAA	180
CATGTATTAA	GTGGAACATT	AGGAGTACCA	GAACATACAT	ATAGAAGTAG	AGTAAATTTG	240
TTTAGTGATA	GATTCATAAA	AGTATTAAAC	TTAGCAAATT	TTACAACAAA	AGATGAAGGA	300
GATTATATGT	GTGAGCTCAG	AGTAAGTGGA	CAAAATCCAA	CAAGTAGTAA	TAAAACAATA	360
AATGTAATAA	GAGATAAATT	AGTAAAATGT	GGAGGAATAA	GTTTATTAGT	ACAAAATACA	420
AGTTGGTTAT	TATTATTATT	ATTAAGTTTA	AGTTTTTTAC	AAGCAACAGA	TTTATAAGT	480
TTATGA						486

- 59 -

## (2) INFORMATION FOR SEQ ID NO:37:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 485 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:37:

ATGAACCCAG TCATCAGCAT CACTCTCCTG CTTTCAGTCT TGCAGATGTC CCGAGGACAG	60
AGGGTGATCA GCCTGACAGC CTGCCTGGTG AACAGAACCT TCGACTGGAC TGCCGTCATG	120
AGAATAACAC CAACTTGCCC ATCCAGCATG AGTTCAGCCT GACCCGAGAG AAGAAGAAGC	180
ACGTGCTGTC AGGCACCCTG GGGGTTCCCG AGCACACTTA CCGCTCCCGC GTCAACCTTT	240
TCAGTGACCG CTTTATCAAG GTCCTTACTC TAGCCAACTT GACCACCAAG GATGAGGGCG	300
ACTACATGTG TGAACCTCGA GTCTCGGGCC AGAATCCCAC AAGCTCCAAT AAAACTATCA	360
ATGTGATCAG AGACAAGCTG GTCAAGTGTG GTGGCATAAG CCTGCTGGTT CAAAACACTT	420
CCTGGCTGCT GCTGCTCCTG CTTTCCCTCT CCTTCCTCCA AGCCACGGAC TTCATTTCTC	480
TGTGA	485

## (2) INFORMATION FOR SEQ ID NO:38:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 33 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

## (xi) SEQUENCE DESCRIPTION: SEQ ID NO:38:

CGCGGGGCTA GCGCAAAGAG TAATAAGTTT AAC	33
--------------------------------------	----

## (2) INFORMATION FOR SEQ ID NO:39:

- (i) SEQUENCE CHARACTERISTICS:  
(A) LENGTH: 28 base pairs  
(B) TYPE: nucleic acid  
(C) STRANDEDNESS: single  
(D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

- 60 -

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:39:

CGCGGATCCC TTGTATTTTG TACTAATA

28

(2) INFORMATION FOR SEQ ID NO:40:

(i) SEQUENCE CHARACTERISTICS:

- (A) LENGTH: 762 base pairs
- (B) TYPE: nucleic acid
- (C) STRANDEDNESS: single
- (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: DNA (genomic)

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:40:

GAATTCACGC GTAAGCTTGC CGCCACCATG GTGAGCAAGG GCGAGGAGCT GTTCACCGGG	60
GTGGTGCCCA TCCTGGTCGA GCTGGACGGC GACGTGAACG GCCACAAGTT CAGCGTGTCC	120
GGCGAGGGCG AGGGCGATGC CACCTACGGC AAGCTGACCC TGAAGTTCAT CTGCACCACC	180
GGCAAGCTGC CCGTGCCCTG GCCCACCCTC GTGACCACCT TCAGCTACGG CGTGCAGTGC	240
TTCAGCCGCT ACCCCGACCA CATGAAGCAG CACGACTTCT TCAAGTCCGC CATGCCCGAA	300
GGCTACGTCC AGGAGCGCAC CATCTTCTTC AAGGACGACG GCAACTACAA GACCCGCGCC	360
GAGGTGAAGT TCGAGGGCGA CACCCTGGTG AACCGCATCG AGCTGAAGGG CATCGACTTC	420
AAGGAGGACG GCAACATCCT GGGGCACAAG CTGGAGTACA ACTACAACAG CCACAACGTC	480
TATATCATGG CCGACAAGCA GAAGAACGGC ATCAAGGTGA ACTTCAAGAT CCGCCACAAC	540
ATCGAGGACG GCAGCGTGCA GCTCGCCGAC CACTACCAGC AGAACACCCC CATCGGCGAC	600
GGCCCCGTGC TGCTGCCCCA CAACCACTAC CTGAGCACCC AGTCCGCCCT GAGCAAAGAC	660
CCCAACGAGA AGCGCGATCA CATGGTCCTG CTGGAGTTCG TGACCGCCGC CGGGATCACT	720
CACGGCATGG ACGAGCTGTA CAAGTAAAGC GGCCGCGGAT CC	762

- 61 -

What is claimed is:

1. A synthetic gene encoding a protein normally expressed in a eukaryotic cell wherein at least one non-preferred or less preferred codon in the natural gene  
5 encoding said protein has been replaced by a preferred codon encoding the same amino acid.
2. The synthetic gene of claim 1 wherein said synthetic gene is capable of expressing said eukaryotic protein at a level which is at least 110% of that  
10 expressed by said natural gene in an in vitro mammalian cell culture system under identical conditions.
3. The synthetic gene of claim 1 wherein said synthetic gene is capable of expressing said eukaryotic protein at a level which is at least 150% of that  
15 expressed by said natural gene in an in vitro cell culture system under identical conditions.
4. The synthetic gene of claim 1 wherein said synthetic gene is capable of expressing said eukaryotic protein at a level which is at least 200% of that  
20 expressed by said natural gene in an in vitro cell culture system under identical conditions.
5. The synthetic gene of claim 1 wherein said synthetic gene is capable of expressing said eukaryotic protein at a level which is at least 500% of that  
25 expressed by said natural gene in an in vitro cell culture system under identical conditions.
6. The synthetic gene of claim 1 wherein said synthetic gene is capable of expressing said eukaryotic protein at a level which is at least ten times that

- 62 -

expressed by said natural gene in an in vitro cell culture system under identical conditions.

7. The synthetic gene of claim 1 wherein at least 10% of the codons in said natural gene are non-preferred  
5 codons.

8. The synthetic gene of claim 8 wherein at least 50% of the codons in said natural gene are non-preferred codons.

9. The synthetic gene of claim 1 wherein at least  
10 50% of the non-preferred codons and less preferred codons present in said natural gene have been replaced by preferred codons.

10. The synthetic gene of claim 1 wherein at least 90% of the non-preferred codons and less preferred  
15 codons present in said natural gene have been replaced by preferred codons.

11. The synthetic gene of claim 1 wherein said protein is green fluorescent protein.

12. A method for preparing a synthetic gene  
20 encoding a protein normally expressed by eukaryotic cells, comprising identifying non-preferred and less-preferred codons in the natural gene encoding said protein and replacing one or more of said non-preferred and less-preferred codons with a preferred codon encoding  
25 the same amino acid as the replaced codon.



Syngpl20mn

1 CTCGAGATCC ATTGTGCTCT AAAGGAGATA CCCGGCCAGA CACCCTCACC  
51 TGCSTGCCCC AGCTGCCCAG GCTGAGGCAA GAGAAGGCCA GAAACCATGC  
101 CCATGGGGTC TGTGCAACCG CTGGCCACCT TGTACCTGCT GGGGATGCTG  
151 GTCCGTTCCG TGCTAGCCAC CGAGAAGCTG TGGGTGACCG TGTACTACCG  
201 CGTGCCCGTG TGAAGGAGG CCACCACCAC CCTGTTCTGC GCCAGCGACG  
251 CCAAGGCGTA CGACCCGAG GTGCACAACG TGTGGGCCAC CCAGGCGTGC  
301 GTGCCCCACG ATCCCAACCC CCAGGAGGTG GAGCTCGTGA ACGTGACCGA  
351 GAACTTCAAC ATGTGGAAGA ACAACATGGT GGAGCAGATG CATGAGGACA  
401 TCATCAGCCT GTGGGACCAG AGCCTGAAGC CCTGCGTGAA GCTGACCCCC  
451 CTGTGCGTGA CCGTGAAGTG CACCGACCTG AGGAACACCA CCAACACCAA  
501 CAACAGCACC GCCAACAACA ACAGCAACAG CGAGGGCACC ATCAAGGGCG  
551 GCGAGATGAA CAACTGCAGC TTCAACATCA CCACCAGCAT CCGCGACAAG  
601 ATGCAGAAGG AGTACGCCCC GCTGTACAAG CTGGATATCG TGAGCATCGA  
651 CAACGACAGC ACCAGCTACC GCCTGATCTC CTGCAACACC AGCGTGATCA  
701 CCCAGGCCTG GCCCAAGATC AGCTTCGAGC CCATCCCCAT CCACTACTGC  
751 GCGCGCGCCG GTTTCGCCAT CCTGAAAGTGC AACGACAAGA AGTTTCAGCGG  
801 CAAGGGCAGC TGCAAGAACC TGAGCACCGT GCAGTGCACC CACGGCATCC  
851 GGCCGGTGGT GAGCACCCAG CTCCTGCTGA ACGGCAGCCT GGCCGAGGAG  
901 GAGGTGGTGA TCCGCAGCGA GAACTTCACC GACAACGCCA AGACCATCAT  
951 CGTGACCTG AATGAGAGCG TGCAGATCAA CTGCACGCGT CCCAACTACA  
1001 ACAAGCGCAA GCGCATCCAC ATCGGCCCCG GCGCGCGCTT CTACACCACC  
1051 AAGAACATCA TCGGCACCAT CCGCCAGGCC CACTGCAACA TCTCTAGAGC  
1101 CAAGTGGAAC GACACCCTGC GCCAGATCGT GAGCAAGCTG AAGGAGCAGT  
1151 TCAAGAACAA GACCATCGTG TTCAACCAGA GCAGCGGCGG CGACCCCGAG  
1201 ATCGTGATGC ACAGCTTCAA CTGCGGCGGC GAATTCCTCT ACTGCAACAC  
1251 CAGCCCCCTG TTCAACAGCA CCTGGAACGG CAACAACACC TGGAAACAAC  
1301 CCACCGGCAG CAACAACAAT ATTACCCTCC AGTGCAAGAT CAAGCAGATC  
1351 ATCAACATGT CGCAGGAGGT GGGCAAGGCC ATGTACGCCC CCCCCATCGA  
1401 GGGCCAGATC CCGTGACGCA GCAACATCAC CCGTCTGCTG CTGACCCCGG  
1451 ACGGCGGCAA GGACACCGAC ACCAACGACA CCGAAATCTT CCGCCCCGGC

FIG 1  
(SHEET 1 OF 4)

2/14

1501 GGGGGCGACA TGGCGGACAA CTGGAGATCT GAGCTGTACA AGTACAAGGT  
1551 GGTGACGATC GAGCCCCTGG GCGTGGCCCC CACCAAGGCC AAGCGCCGCG  
1601 TGGTGCAGCG CAGAGAAGCGC TAAAGCGGCC GC (SEQ ID NO:34)

FIG 1  
(SHEET 2 OF 4)

Syn gp160 mRNA

1 ACCGAGAAGC TGTGGGTGAC CTTGTACTAC GCGGTGCGCG TGTGGAAGGA  
31 GGGCACCACC ACCCTTTTCT GCGCCAGCGA GCGCAAGGCG TACGACACCG  
61 AUGTGCACAA CTTGTGGGCG ACCCAGGCGT GCGTGCCAC CGACCCCAAC  
91 GGGCAGGAGG TGGAGCTCTT GAACGTGACC GAGAACTTCA ACATGTGGAA  
121 GAACAACATG CTGGAGCAGA TGCATGAGGA CATCATCAGC CTGTGGGACC  
151 AGAGCCTGAA GCGCTGCGTG AAGCTGACCG CCGTGTGCGT GACCCCTCAAC  
181 TGCACCGAGC TTAGGAACAC CACCAACACC AACAACAGCA CCGCCCAACA  
211 CAACATCAAC AGCGAGGGCA CCATCAAGGG CCGCGAGATG AAGAACTGCA  
241 GCTTCAACAT CACCACCAGC ATCCGCGACA AGATCCAGAA GGATACCGCC  
271 CTGCTGTACA AGCTGGATAT CCGGAGCATC CACAACGACA GCAACAGCTA  
301 CCGCTGTATC TCGTCAACA CCAAGCTGAT CACCGAGGCG TCGCCCAAGA  
331 TCAGCTTCCA GCGCATCGCG ATCACTACT GCGCGCGCGC CGGCTTCCGC  
361 ATCTGAACT GCAACGACA GAAGTTCAGC GCGAAGGGCA CCGCAAGAA  
391 CCGTACCACC CTGCACTGCA CCGACCGCAT CCGCGCGGTG GTGAGCAGCC  
421 ACCTCTCTCT GAACGGCAGC CTGCGCGAGG AGGAGGTGCT GATCGGCGC  
451 GAGAACTTCA CCGACAACCG CAAGACCATC ATCTGCGACC TGAATGAGAG  
481 CCGTGCAGATC AACTGCACCG GTCCCAACTA CAACAAGCGC AAGCGCATCC  
511 ACATCGGCGC CCGCGCGCGC TTCTACACCA CCAAGAACAT CATCGGCACC  
541 ATCCGCGAGC CCGACTGCAA CATCTGTAGA GCGAAGTGGG ACGACACCTT  
571 GCGCCAGATC GTGAGCAAGC TGAAGGAGCA GTTCAAGAAC AAGACCATCC  
601 TGTTCACCA GAGCAGCGCG GCGGACCGCG AGATCGTGAT GCACAGCTTC  
631 AACTGCGGCG GCGAATTCTT CTACTGCAAC ACCAGCGCGC TGTTCACAG  
661 CACCTGGAAC GCGAACAACA CCGGAACAA CACCAACGCG AGCAACAACA  
691 ATATTACCCT CCAGTGCAAG ATCAAGCAGA TCATCAACAT GTGGCAGGAG  
721 GTGGGCAAGG CCATGTACCG CCGCGCGATC GAGGCGCAGA TCGGCTGAG  
751 CAGCAACATC ACCCGTCTGC TCGTGACCG CGACGCGCGC AAGGACACCG  
781 ACACCAACGA CACCGAAATC TCGCGCGCGC GCGCGCGCGA CATCGCGGAC  
811 AACTGGAGAT CTGAGCTGTA CAAGTACAAG GTGGTGACGA TCGAGCGCGT  
841 CCGCGTGGCG CCGACCAAGG CCAAGCGCG CCGTGTGCGC CCGGAGAAGC

FIG. 1  
(SHEET 3 OF 4)

1451 GGGCCGCCAT CCGCGCCCTG TTCCTGGGCT TCCTGGGGGC GCGGGGCAGC  
1501 ACCATGGGGG CCGCCAGCGT GACCCTGACC GTGCAGGCCC GCCTGCTCCT  
1551 GAGCGGCATC GTGCAGCAGC AGAACAACCT CCTCCGCGCC ATCGAGGCCC  
1601 AGCAGCATAT GTTCCAGCTC ACCGTGTGGG GCATCAAGCA GCTCCAGGCC  
1651 CGCGTGCTGG CCGTGGAGCG CTACCTGAAG GACCAGCAGC TCCTGGGCTT  
1701 CTGGGGCTGC TCCGGCAAGC TGATCTGCAC CACACCGTA CCTGGAACG  
1751 CCTCCTGGAG CAACAAGAGC CTGGACGACA TCTGGAACAA CATGACCTGG  
1801 ATGCAGTGGG AGCGCGAGAT CGATAACTAC ACCAGCCTGA TCTACAGCCT  
1851 GCTGGAGAAG AGCCAGACCC AGCAGGAGAA GAACGAGCAG GAGCTGCTGG  
1901 AGCTGGACAA CTGGGCGAGC CTGTGGAACG GGTTCGACAT CACCAACTGG  
1951 CTGTGGTACA TAAAAATCTT CATCATGATT GTGGGCGGCC TGGTGGGCTT  
2001 CCGCATCGTG TTGGCCGTGC TGAGCATCGT GAACCGCGTG CCGCAGGGCT  
2051 ACAGCCCCCT GAGCCTCCAG ACCCGGCCCC CCGTGGCGCG CGGGCCCCGAC  
2101 CGCCCCGAGG CCAATCGAGGA GGAGGGCGGC GAGCGCGACC GCGACACCAG  
2151 CGGCAGGCTC GTGCACGGCT TCCTGGCGAT CATCTGGGTG GACCTCCGCA  
2201 GCCTGTTCTT GTTCAGCTAC CACCACCGCG ACCTGCTGCT GATCGCCGCC  
2251 CGCATCGTGG AACTCCTAGG CCGCCGCGGC TGGGAGGTGC TGAAGTACTG  
2301 GTGGAACCTC CTCCAGTATT GGAGCCAGGA GCTGAAGTCC AGCGCCGTGA  
2351 GCCTGCTGAA CGCCACCGCC ATCGCCGTGG CCGAGGGCAC CGACCGCGTG  
2401 ATCGAGGTGC TCCAGAGGCC CGGGAGGGCG ATCTGCACA TCCCCACCCG  
2451 CATCCGCCAG CGGCTCGAGA GGGCGCTGCT G (SEQ ID NO:35)

FIG. 1

(SHEET 4 OF 4)

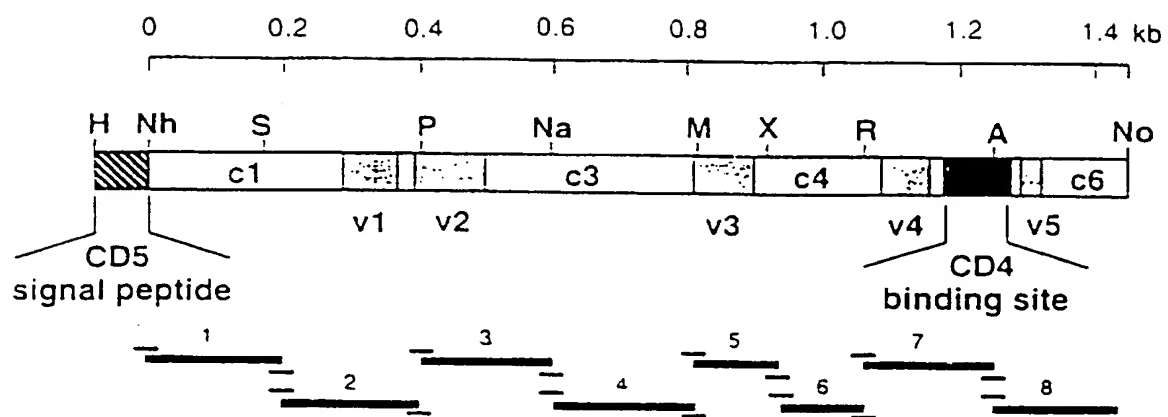


FIGURE 2

6/14

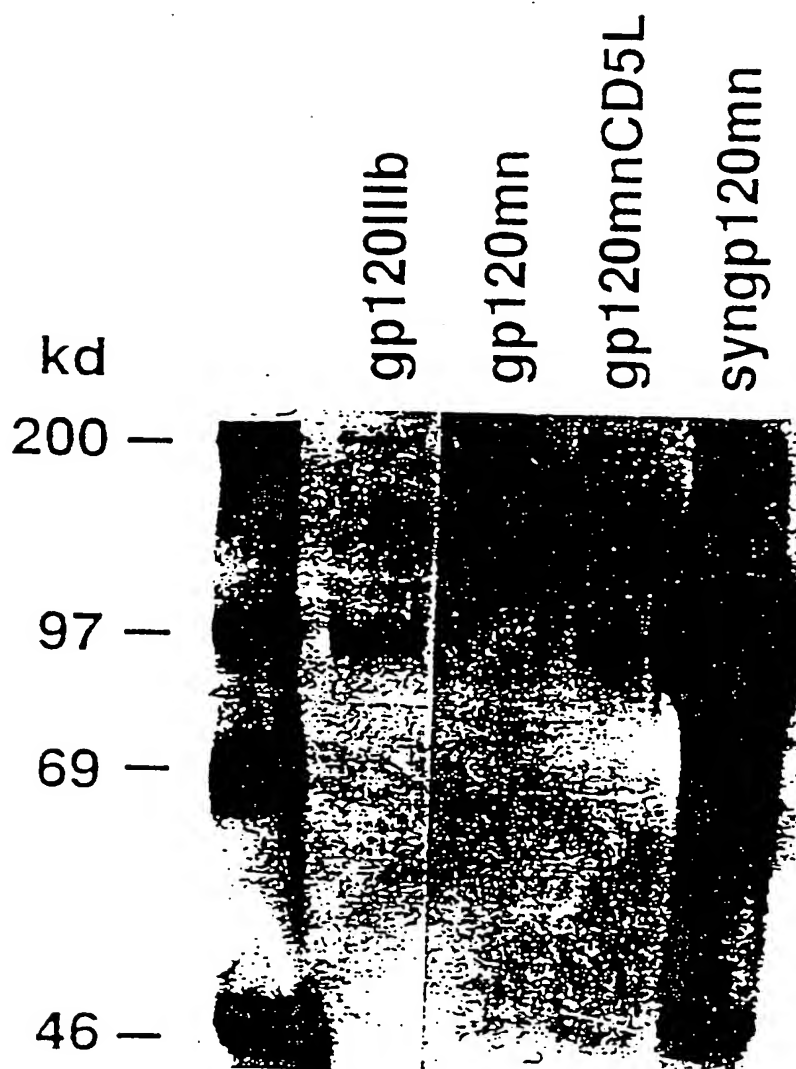


FIGURE 3

7/14

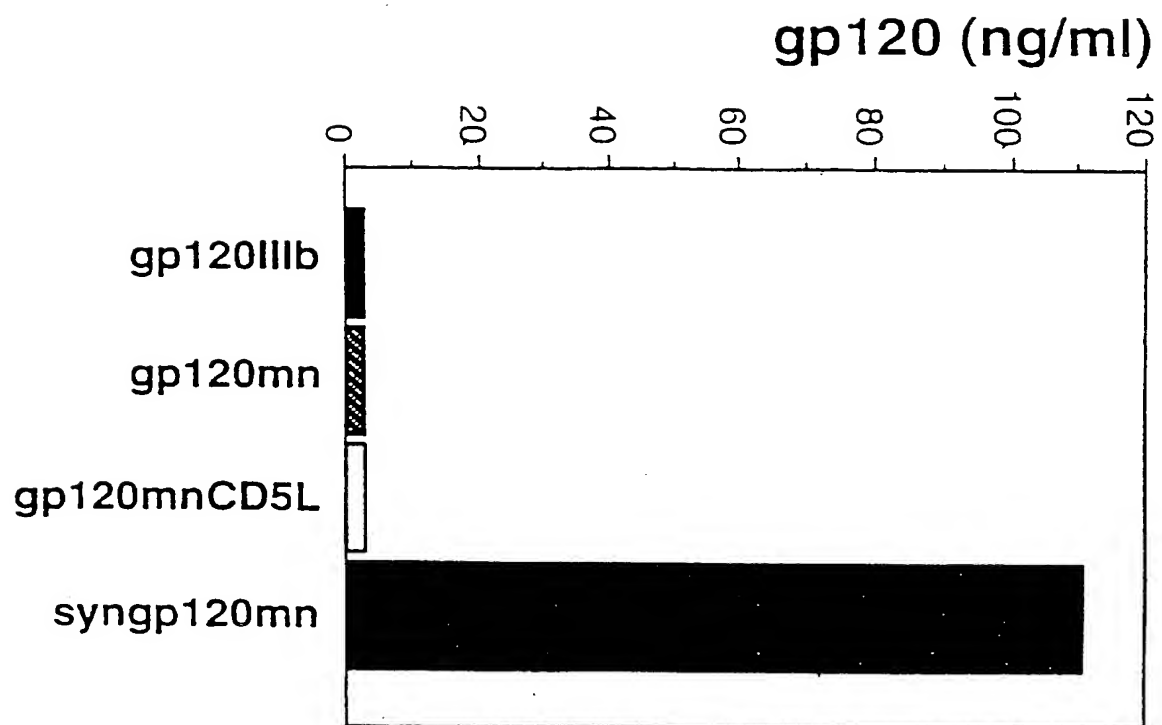


FIGURE 4

8/14

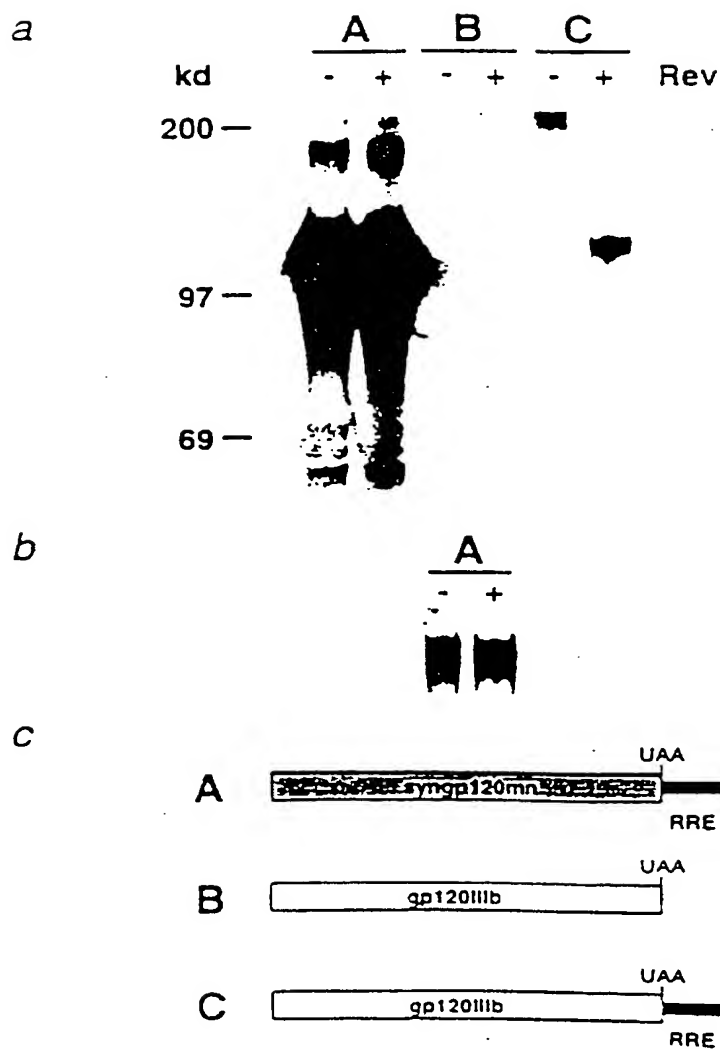


FIGURE 5



9/14

```

      M   N   P   V   I   S   I   T   L   L   L   S   V   L   Q   M   S   R   G   Q
(SEQ ID NO:36) env → atg aat cca gta ata agt ata aca tta tta agt gta tta caa atg agt aga gga caa
(SEQ ID NO:37) wt → atg aac cca gtc atc agc atc act ctc ctg ctt tca gtc ttg cag atg tcc cga gga cag

      R   V   I   S   L   T   A   C   L   V   N   Q   N   L   R   L   D   C   R   H
env   aga gta ata agt tta aca gca tgt tta gta aat caa aat ttg aga tta gat tgt aga cat
wt   agg gtg atc agc ctg aca gcc tgc ctg gtg aa cag aac ctt cga ctg gac tgc cgt cat

      E   N   N   T   N   L   P   I   Q   H   E   F   S   L   T   R   E   K   K   K
env   gaa aat aat aca cct ttg cca ata caa cat gaa ttt tca tta acg cgt gaa aaa aaa
wt   gag aat aac acc aac ttg ccc atc cag cat gag ttc agc ctg acc cga gag aag aag aag

      H   V   L   S   G   T   L   G   V   P   E   H   T   Y   R   S   R   V   N   L
env   cat gta tta agt gga aca tta gga gta cca gaa cat aca tat aga agt aga gta aat ttg
wt   cac gtg ctg tca ggc acc ctg ggt gtt ccc gag cac act tac cgc tcc cgc gtc aac ctt

      F   S   D   R   F   I   K   V   L   T   L   A   N   F   T   K   D   E   G
env   ttt agt gat aga ttc ata aaa gta tta aca tta gca aat ttt aca aca aaa gat gaa gga
wt   ttc agt gac cgc ttt atc aag gtc ctt act cta gcc aac ttc acc acc aag gat gag ggc

      D   Y   M   C   E   L   R   V   S   G   Q   N   P   T   S   S   N   K   T   I
env   gat tat atg tgt gag ctc aga gta agt gga caa aat cca aca agt agt aat aaa aca ata
wt   gac tac atg tgt gaa ctt cga gtc tgc ggc cag aat ccc aca agc tcc aat aaa act atc

      N   V   I   R   D   K   L   V   K   C   G   I   S   L   L   V   Q   N   T
env   aat gta ata aga gat aaa tta gta aaa tgt gga gga ata agt tta tta gta caa aat aca
wt   aat gtg atc aga gac aag ctg gtc aag tgt ggt ggc ata agc ctg gtt cca aac act

      S   W   L   L   L   L   L   L   S   L   S   F   L   Q   A   T   D   F   I   S
env   agt tgg tta tta tta tta tta agt tta agt ttt tta caa gca aca gat ttt ata agt
wt   tcc tgg ctg ctg ctg ctg ctc tcc ctc tcc ctc caa gcc acg gac ttc att tct

      L   *
env   tta tga
wt   ctg tga

```

FIGURE 6

10/14

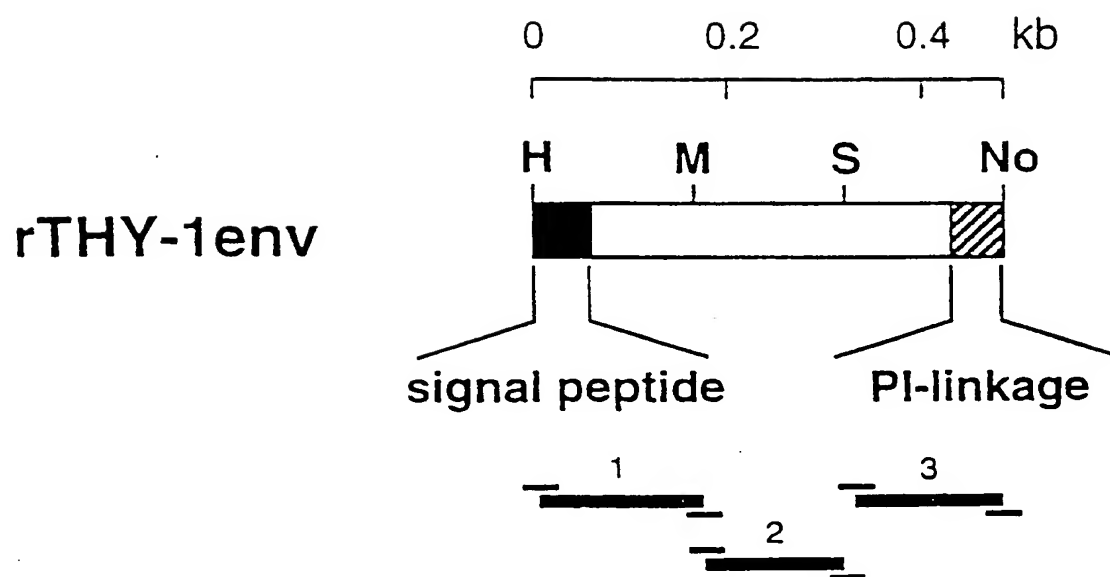


FIGURE 7

11/14

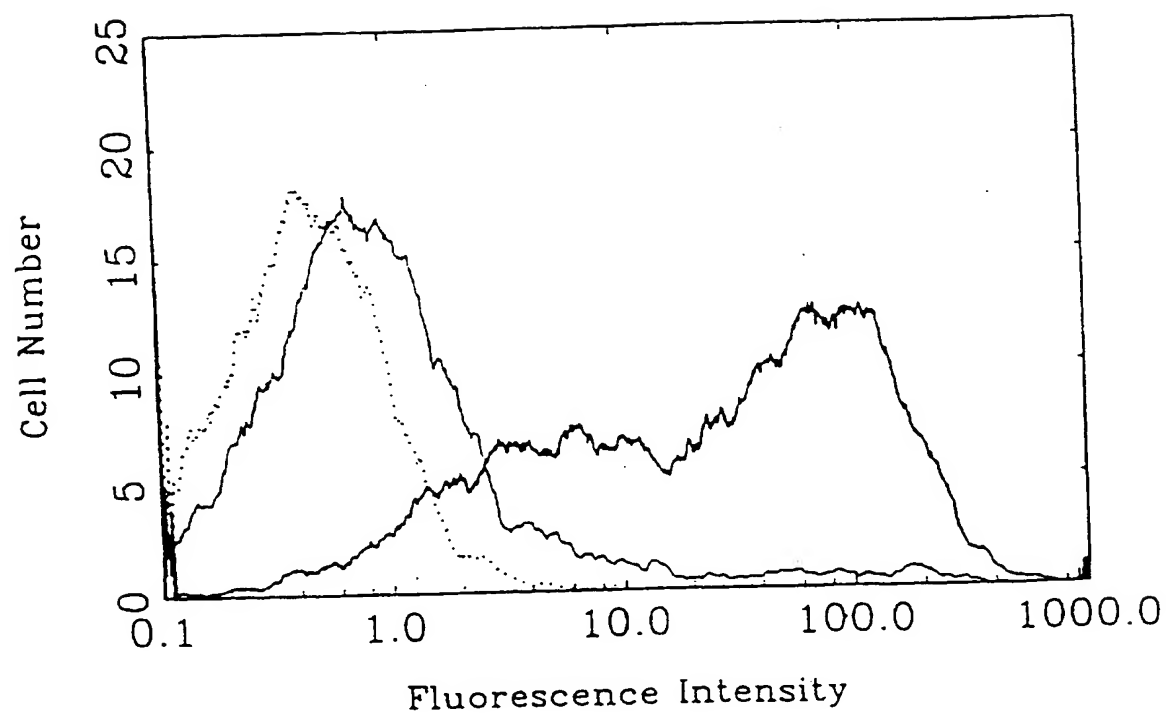


FIGURE 8

12/14

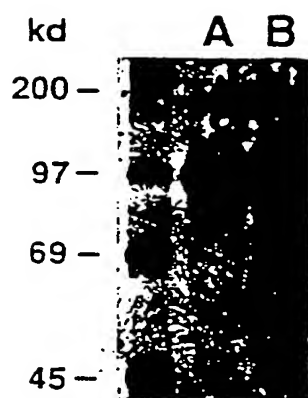
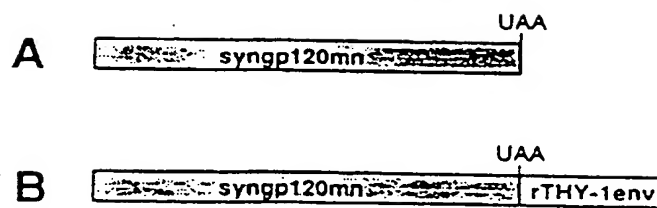
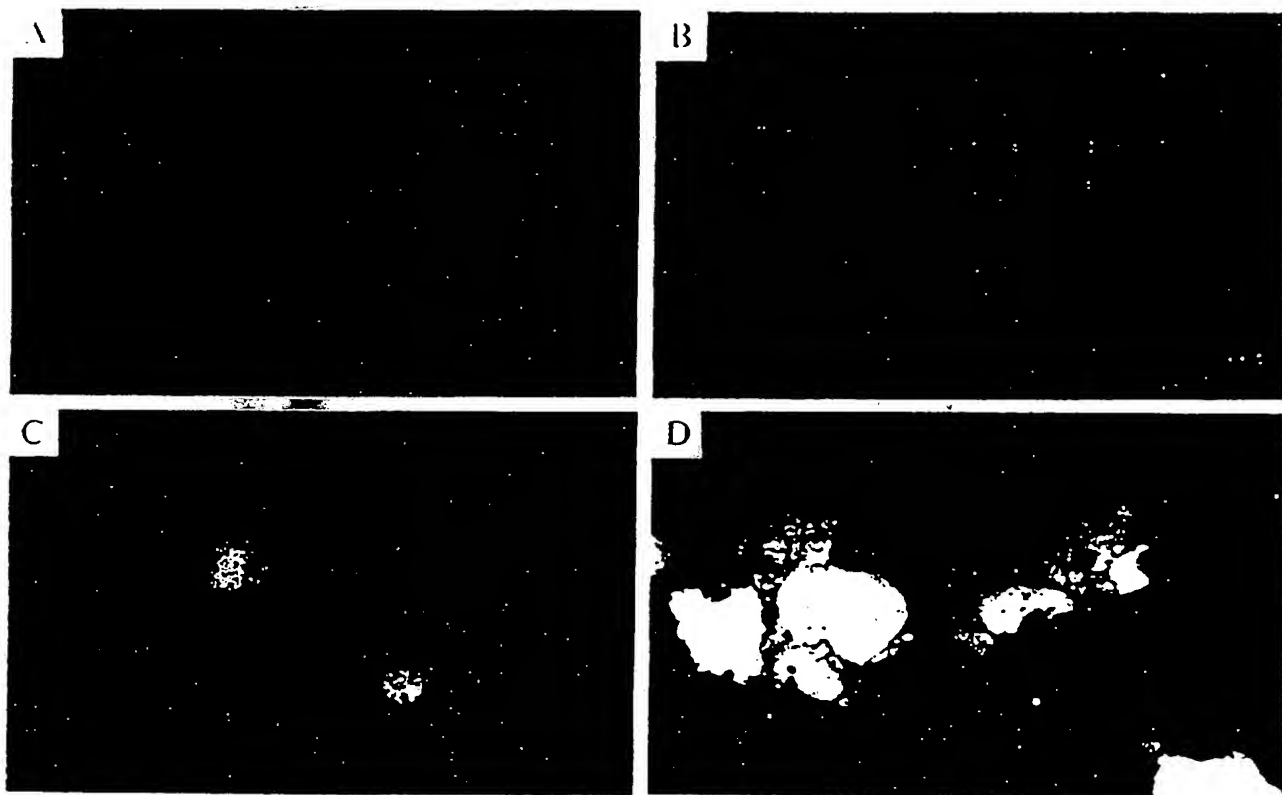
*a**b*

FIGURE 9

FIG. 10



1 GAATTCACGC GTAAGCTTGC CGCCACCATG GTGAGCAAGG GCGAGGAGCT  
51 GTTACACGGG GTGGTGCCCA TCCTGGTCGA GCTGGACGGC GACGTGAACG  
101 GCCACAAGTT CAGCGTGTCC GGCAGGGGCG AGGGCGATGC CACCTACGGC  
151 AAGCTGACCC TGAAGTTCAT CTGCACCACC GGCAAGCTGC CCGTGCCCTG  
201 GCCCACCCTC GTGACCACCT TCAGCTACGG CGTGCACTGC TTCAGCCGCT  
251 ACCCCGACCA CATGAAGCAG CACGACTTCT TCAAGTCCGC CATGCCCCGAA  
301 GGCTACGTCC AGGAGCGCAC CATCTTCTTC AAGGACGACG GCAACTACAA  
351 GACCCGCGCC GAGGTGAAGT TCGAGGGCGA CACCCTGGTG AACCGCATCG  
401 AGCTGAAGGG CATCGACTTC AAGGAGGACG GCAACATCCT GGGGCACAAG  
451 CTGGAGTACA ACTACAACAG CCACAACGTC TATATCATGG CCGACAAGCA  
501 GAAGAACGGC ATCAAGGTGA ACTTCAAGAT CCGCCACAAC ATCGAGGACG  
551 GCAGCGTGCA GCTCGCCGAC CACTACCAGC AGAACACCCC CATCGGCGAC  
601 GGCCCCGTGC TGCTGCCCCGA CAACCACTAC CTGAGCACCC AGTCCGCCCT  
651 GAGCAAAGAC CCCAACGAGA AGCGCGATCA CATGGTCCTG CTGGAGTTCG  
701 TGACCGCCGC CGGGATCACT CACGGCATGG ACGAGCTGTA CAAGTAAAGC  
751 GGCCGCGGAT CC (SEQ ID NO: 40)

FIG. 11

# INTERNATIONAL SEARCH REPORT

International application No.  
PCT/US96/15088

## A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C07H 21/00, 21/04

US CL : 536/23.1, 23.5

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 536/23.1, 23.5

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

Dialog, Medline, Biosis, Embase, Scisearch, WPIDS, APS

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	HOLLER et al. HIV1 Integrase Expressed in Escherichia coli From a Synthetic Gene. Gene. 1993, Vol.136, pages 323-328, especially pages 323-327.	1-10, 12
X	SCORER et al. The Intracellular Production and Secretion of HIV-1 Envelope Protein in the Methylophilic Yeast Pichia pastoris. Gene. 1993, Vol.136, pages 111-119, especially pages 111-118.	1-10, 12
X	HERNAN et al. Human Hemoglobin Expression in Escherichia coli: Importance of Optimal Codon Usage. Biochemistry. 1992, Vol.31, pages 8619-8628, especially pages 8619-8627.	1-10, 12

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:	*T	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
*A* document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
*E* earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*G*	document member of the same patent family
*O* document referring to an oral disclosure, use, exhibition or other means		
*P* document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search

26 NOVEMBER 1996

Date of mailing of the international search report

23 JAN 1997

Name and mailing address of the ISA/US  
Commissioner of Patents and Trademarks  
Box PCT  
Washington, D.C. 20231

Facsimile No. (703) 305-3230

Authorized officer

ENRIQUE D. LONGTON

Telephone No. (703) 308-0196

## INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/15088

## C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X ---- Y	WILLIAMS et al. Design, Synthesis and Expression of a Human Interleukin-2 Gene Incorporating the Codon Usage Bias Found in Highly Expressed Escherichia coli Genes. Nucleic Acids Research. 1988, Vol.16, No.22, pages 10453-10467, especially pages 10453-10466.	1-10, 12 ---- 11
X	RANGWALA et al. High-Level Production of Active HIV-1 Protease In Escherichia coli. Gene. 1992, Vol.122, pages 263-269, especially pages 263-268.	1-10, 12
P, X	US 5,464,774 A (BAIRD et al.) 07 November 1995 (07/11/95), see entire document, especially insert at top of columns 13 and 14; column 7, lines 27-51.	1-10, 12
Y	INOUE et al. Aequorea Green Fluorescent Protein Expression of the Gene and Fluorescence Characteristics of the Recombinant Protein. FEBS Letters. 1994, Vol.341, pages 277-280, especially pages 277-279.	11

Form PCT/ISA/210 (continuation of second sheet)(July 1992)\*